AD________________

Award Number: W81XWH-07-1-0242

TITLE: Infrared Spectroscopic Imaging for Prostate Pathology Practice

PRINCIPAL INVESTIGATOR: Rohit Bhargava, Ph.D.

CONTRACTING ORGANIZATION: University of Illinois
Champaign, IL 61820

REPORT DATE: April 2011

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE
1 Apr 2011

2. REPORT TYPE
Final

3. DATES COVERED
15 FEB 2007 - 31 MAR 2011

4. TITLE AND SUBTITLE

Infrared Spectroscopic Imaging for Prostate Pathology Practice

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-07-1-0242

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Rohit Bhargava, Ph.D.

E-Mail: rxb@illinois.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Illinois
Champaign, IL 61820

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The report summarizes progress towards using Fourier transform infrared spectroscopic imaging for prostate pathology in year 3 of a 3 year award from the PCRP. The aim of the work is to enable histopathologic recognition without the use of human input or stains.

15. SUBJECT TERMS
Spectroscopy, prostate, histopathology, cancer, optimization, optical imaging

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** USAMRMC
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	223	**19b. TELEPHONE NUMBER** *(include area code)*

Table of Contents

Page

Introduction………………………………………………………….……….….. 3

Body………………………………………………………………………………….. 4

Key Research Accomplishments…………………………………….…….. 44

Reportable Outcomes…………………………………………………………… 44

Conclusion……………………………………………………………………… 49

References………………………………………………………………………. 49

Appendices……………………………………………………………………… 51

Introduction

Prostate cancer accounts for one-third of noncutaneous cancers diagnosed in US men,[1] is a leading cause of cancer-related death and is, appropriately, the subject of heightened public awareness and widespread screening. If prostate-specific antigen (PSA)[2] or digital rectal screens are abnormal,[3] a biopsy is considered to detect or rule out cancer. Pathologic status of biopsied tissue forms the definitive diagnosis for prostate cancer and constitutes an important cornerstone of therapy and prognosis.[4] There is, hence, a need to add useful information to diagnoses and to introduce new technologies that allow efficient analyses of cancer to focus limited healthcare resources. For the reasons underlined above, there is an urgent need for high-throughput, automated and objective pathology tools. Our general hypothesis is that these requirements are satisfied through innovative spectroscopic imaging approaches that are compatible with, and add substantially to, current pathology practice. Hence, the overall aim of this project is to demonstrate the utility of novel Fourier transform infrared (FTIR) spectroscopy-based, computer-aided diagnoses for prostate cancer and develop the required microscopy and software tools to enable its application.

FTIR spectroscopic imaging is a new technique that combines the spatial specificity of optical microscopy and the biochemical content of spectroscopy.[5] As opposed to thermal infrared imaging, FTIR imaging measures the absorption properties of tissue through a spectrum consisting of (typically) 1024 to 2048 wavelength elements per pixel.[6] Since mid-IR (2-12 µm wavelength) spectra reflect the molecular composition of the tissue, image contrast arises from differences in endogenous chemical species. As opposed to visible microscopy of stained tissue that requires a human eye to detect changes, numerical computation is required to extract information from IR spectra of unstained tissue. Extracted information, based on a computer algorithm, is inherently objective and automated. Recent work has demonstrated that these determinations are also accurate and reproducible in large patient populations.[7] Hence, we focused, in the first year of this project, on demonstrating that the laboratory results could be optimized using novel approaches to fast imaging. This is a critical step, since we propose next to analyze 375 radical prostatectomy samples. We have been able to optimize data acquisition parameters and develop a novel algorithm for processing data that enables almost 50-fold faster imaging. Briefly, the idea behind the process is illustrated in Fig 1. In this performance period, we sought to use acquired data to establish the use of IR imaging for validating cancer diagnosis (task 2), develop a calibration and prediction model for grading and perform extensive validation (task 2). Finally, we sought to develop a mathematical framework to relate disparate pieces of information to outcome (task 3).

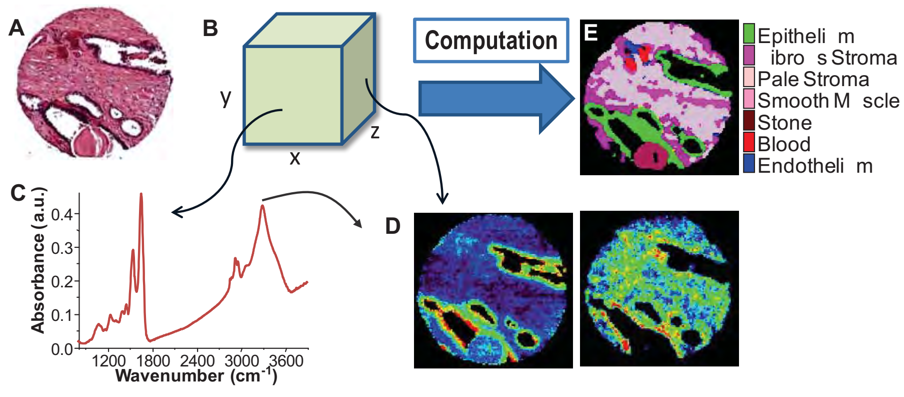


Figure 1. (A) Conventional imaging in pathology requires dyes and a human to recognize cells. In chemical imaging data cubes (B), both a spectrum at any pixel (C) and the spatial distribution of any spectral feature can be seen. e.g. in (D) nucleic acids (left, at ~1080 cm^{-1}), and collagen specific (right, at ~ 1245 cm^{-1}) Computational tools can then convert chemical imaging data to knowledge used in pathology (E).

Body

Specific activities and tasks as per statement of work during this performance period are described below. Details of performance for the past years periods are given in the past annual reports which is attached for quick reference of the reviewers. :

***Task 1*. Perform infrared spectroscopic imaging on prostate biopsy specimens**

Goal: Data will be acquired from samples identified in Task 2, sub-task a. 4 cm^(-1) spectral resolution data, imaging ~6 micrometer of sample per pixel will be acquired with a signal to noise ratio of greater than 1000:1. At least 375 samples will be imaged to provide as estimated 40 million spectra. Data will continuously be available for analysis in this period. (Months 8-18)

Activities: **Activities**: A focal plane array (FPA) detector was interfaced to an infrared interferometer and microscope to record high-throughput spectroscopic imaging data. A rapid-scanning FTIR imaging system that can image more than 16,000 spectra per second was available. The system, however, provided low signal to noise ratio (SNR) data. In increasing the SNR of data acquired, there are typically hardware or experimental approaches. It is prohibitively expensive to procure new hardware. Hence, typically, the approach has been to increase SNR by averaging successively acquired images. The benefits in SNR are $\sqrt{n}$, where n is the numbers of averaged spectral data cubes. Hence, we focused next on developing post-processing methods, as detailed next.

Goal: Develop a route to mathematically transform data to eliminate noise and yield high quality data. A custom algorithm will be developed in which the covariance matrix is employed to first perform a factor analysis equivalent operation followed by image separation from noise and re-transformation. Software to automatically correct data will be available. (Months 2-6)
Activities: The methodology was developed and is demonstrated to show a 50-fold improvement in SNR. Results are reported in publication to *Analyst* and were presented at 2 conferences.

Our approach was the following: The first and simplest approach to higher fidelity imaging required co-adding a large number of array detector snapshots of the same scene, resulted in long dwell times of the mirror at every optical retardation[8]. We operated the interferometer in step-scan mode and wrote custom software to analyze the data. The advantages of this frame co-addition process were limited due to the noise characteristics of the detector. Hence, an optimal combination of frame co-addition and repeated scanning was implemented, as previously proposed[9]. Though these methods make the best use of the available hardware, they unfortunately, require large increases in data acquisition time as the SNR reduction scales less than linearly with the acquisition time. In order to obtain high SNR data using acquisition-side approaches, the trade-off with respect to time is unavoidable. Such a trade-off limits the possible applications of FT-IR imaging as a routine microscopic analysis tool in prostate cancer.

For a finite data acquisition time, other schemes to extract low noise information are available[10] but these methods neglect the image as a whole and result in loss of image fidelity. While we implemented these schemes here, it was clear that structural fidelity of the tissue image was being affected. Hence, we turned our attention to another alternative to hardware improvement or co-addition schemes for high fidelity imaging. This approach is the use of mathematical noise reduction techniques. For example, a procedure based on the Minimum Noise Fraction (MNF) transform was adopted from the satellite and airborne imaging community[11]. With rapid development of powerful computers and increased storage capacities, using computation to enhance instrument performance is becoming an attractive option. Using chemometric methods to enhance acquired FT-IR imaging data has been a relatively recent development. A convenient approach is to use an Eigenvalue decomposition of the data using a forward transform, e.g. PCA. After selecting eigenimages with sufficient SNR, the selected data are inverse transformed to yield the entire dataset with lower noise content. This approach was used[12] to examine phase compositions by enhancing contrast between different regions. PCA reorders data in decreasing order of variance.

A similar technique called MNF transform was proposed[13] to re-order image data in decreasing order of SNR. A modified version[14] of this transform has been shown to improve image fidelity and achieve better noise reduction than PCA, for example.

Mathematical transform techniques for noise reduction generally utilize the fact that noise in uncorrelated whereas spectra (signals) have a fairly high degree of correlation. In the transform domain, the signal is primarily restricted to a few factors where as the noise is spread across all factors. We use the term 'factors' to refer to images of eigenvalues in the transform domain. Noise reduction can be achieved by retaining factors corresponding to high signal content, removing factors predominantly corresponding to noise and computing the inverse transform. Identifying factors corresponding to high signal content is an important step in the noise reduction process.

The identification of factors to include is invariably a manual process and is the key impediment to routine application of these methods for noise reduction. First, the manual selection will vary from practitioner to practitioner, leading to variance in the results obtained from the same data set. The scientific conclusions or confidence in results, hence, may vary in an unpredictable manner. Second, the need to examine every eigenvalue image (or, at least, a large set of images) is time-consuming. The decision to exclude or include images with questionable content is especially difficult and requires significant time as some quantitative guidance is often used. For example, we have used comparisons of values from sample and sample-less regions. These two factors are a key barrier in the use of these post-processing techniques for enhancing IR imaging data.

There are many dimension reduction and noise reduction schemes proposed[15,16]. Many of these methods[15,17,18] choose all factors before a certain cut off (k) determined based on predefined criteria. However, the assumption that all of the first k factors are important is questionable. The MNF approach was specifically developed to overcome the observation that the first k factors in PCA were not always optimal. Other methods[16,19] can be computationally expensive or do not utilize some of the features of the data in factors.

A general criticism of these methods is that they do not explicitly account for the spatial and spectral information in the data. For example, PCA separates features in the spatial domain by accounting for variance in the scene. The variance may arise from the data, sensor or may be an artifact. Similarly, the signal in the re-ordering of MNF factors is assumed to be features in the image but could come from factors other than the sample of interest. For example, Figure 1 shows the 4th, 8th, 12th and 19th MNF factor for FT-IR data from a breast tissue sample. The 4th MNF factor shows interesting tissue structural features. Although the 8th factor has higher SNR compared to the 12th or 19th factor, the 12th and 19th factors contain relatively more features of interest. We would include the 12th and 19th factors but not the 8th in a noise reduction scheme involving MNF transform. The 8th factor likely arises from illumination or water vapor differences and not from the sample itself.

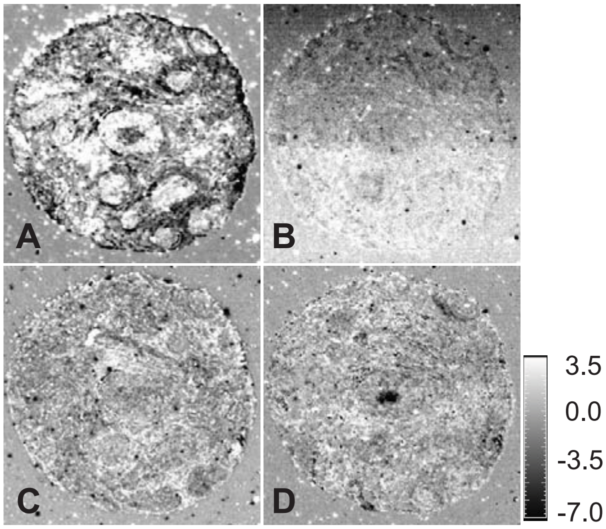


Figure 2. (A) 4th MNF Factor (Tissue structural features visible) (B) 8th MNF factor (C) 12th MNF factor (D) 19th MNF factor. The 8th factor has less structural features compared to 12th or 19th factor.

Hence, we proposed a factor selection algorithm that selects factors based on structural features in a quantitative manner. Although we illustrate the utility of the proposed algorithm for tissue

FT-IR data, the technique is more general and can be applied to any other data in which structures in images are well described by edges. We could also use the proposed factor selection algorithm with other transform techniques like PCA for example. A generalization of the MNF transform has been proposed by[20]. However, we did not observe the kind of distortion described in 20 in our data and therefore did not find the need to use the generalized MNF. We demonstrate the efficacy of this automated SNR enhancement by applying the process to breast tissue data. The effects of SNR are quantitatively measured by the accuracy of classifying tissue.

Over 5 million spectra have been acquired from approx. 475 samples using 4 cm^{-1} resolution over the 7200-720 cm^{-1} range and 6.25 micron on a side per pixel. Data handling and analysis is on-going. The data were acquired using a tissue microarray with no restrictions on age or prior PSA reading. The archiving and record keeping for such data sets became a challenge. Hence, we developed data handling tools to both maintain a database of properties as well as visualize the data in a microarray format. For example, one acquired data set is shown below in Figure 3.

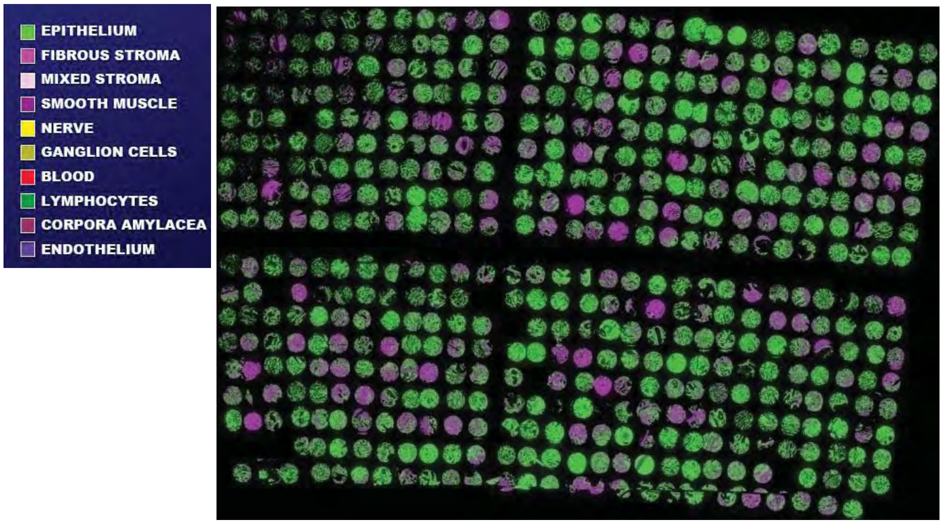


Figure 3. Approximately 475 viable samples for further analysis acquired by FT-IR imaging and classified as per optimized protocols developed previously in this project.

A second set of 460 samples were also acquired for validation studies. This large scale data acquisition has never been previously reported and is a direct result of the optimizations accomplished in year 1 of this project. Corresponding to each sample in the tissue array above, we have developed a database to store information for the patient, including age, PSA values at the time of diagnosis, Gleason grade and stage on diagnosis as well as outcome.

As per previous studies in year 1, we determined that there was a need to acquire data of a signal to noise ratio (SNR) of at least 1000:1 (or, 30 dB). One outstanding question is how to predict the required SNR for any classification task. This is a major issue in which no useful guidance was available in the literature. In observing the data from many samples, it became clear that new tools were needed to visualize diversity and usefulness of particular samples. In particular, one key element of the protocol depends on a quality check. If contaminations exist in samples or the sample does not belong to a population that is similar to the one that was used to construct a calibration of the data, then the sample will clearly lead to incorrect results. Such a sample must be flagged during quality control but there was no obvious means to do so. Hence, we developed a new visualization system for spectrum wide analysis of the data.

First, we recall that not every point in the spectrum is actually useful in calibration or prediction. The data are reduced to a potential set of descriptors, termed metrics, which are peak height ratios, areas, positions or even spatial indices. Only a few of these metrics are useful in calibration, and consequently, in predicting histopathology. Hence, we employ the visualization only for a set of metrics. A view of the developed software and typical plot resulting from the analysis is shown in Figure 4.

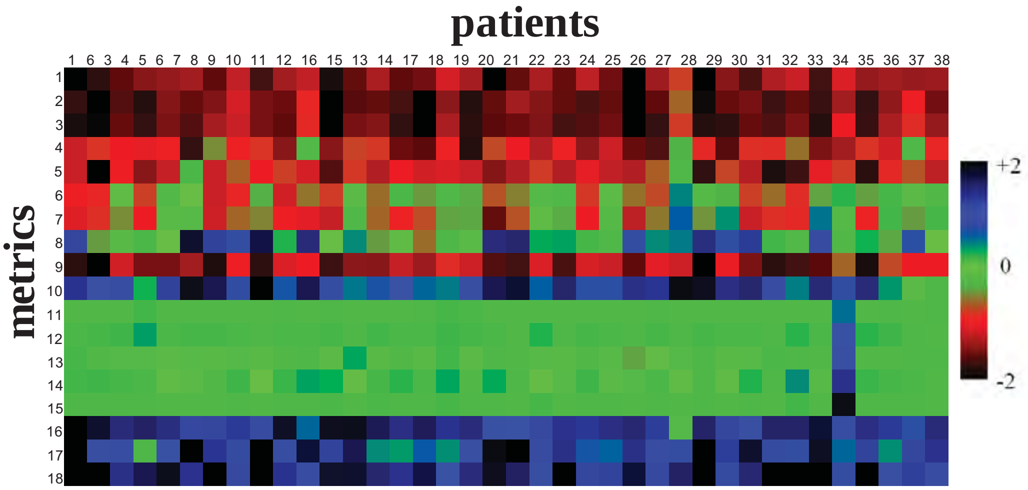


Figure 4. A Representation of metric-patient data to determine quality and consistency in large scale data analysis. Many representations are possible, including the one shown here. Here, the value of (μ_1-μ_2)/σ for each metric is represented, where μ_1 is the mean of epithelium pixels for one patient for a particular metric and μ_2 is the mean of stroma pixels for one patient for a particular metric whereas σ is the standard deviation of the entire metric. Hence, (μ_1-μ_2)/σ is a measure of classification potential in separating epithelium from stroma. Patient no. 34 can be seen to have outlier values that must be investigated in detail so as not to become a confounding variable.

Task 2. Analyze spectroscopic imaging data for biochemical markers of tumor and develop numerical algorithms for grading cancer

Goal: Develop algorithm for malignancy recognition. Models will be constructed and optimized using Genetic Algorithms operating on identified metrics. Models will be tested and validated using ROC curves with pathologist marking as the ground truth. A protocol for segmenting benign from atypical condition will be available. (Months 11-18) Three specific aims from the statement of work (SOW) are:

a. Identify samples to be imaged (Months 1-3) by examining stained slides
b. Obtain unstained samples to be imaged and define regions for calibration and validation (Months 4-7)
c. Perform histologic identification on prostate samples and validate
d. Reduce spectral metrics to those useful in identifying atypia (Months 8-12)

e. Develop protocols and validate distinction between benign-appearing and atypical tissue (Months 12-18)
f. Develop calibration for predicting cancer grade (Months 18-22)
g. Develop protocols and validate Gleason grading of tumor (Months 18-27)

Activities:

Goal: Data acquisition and treatment protocol will be optimized and feedback loop implemented. Image sets will be acquired at low averaging and extensive averaging conditions to verify performance and optimize algorithm. A validated protocol for collecting data will be available. (Months 5-7)
Activities: Data were acquired and experimental conditions were optimized to help determine the operating points for prostate histology. Briefly, the spectral resolution was not found to be important unless coarse resolution was obtained. SNR was found to be crucial and a plot of the SNR versus the classification accuracy yielded the optimal operating point. Results are summarized in a peer-reviewed manuscript[21] and the methodology is described in a review paper. Results were presented at three different meetings.

A single button operation is implemented in our software that now pre-processes data and adjusts for appropriate SNR. A second step can then classify the resulting data into histologically correct classes.

Goal: Data will be acquired from samples identified in Task 2, sub-task a. 4 cm^(-1) spectral resolution data, imaging ~6 micrometer of sample per pixel will be acquired with a signal to noise ratio of greater than 1000:1. At least 375 samples will be imaged to provide as estimated 40 million spectra. Data will continuously be available for analysis in this period. (Months 8-18)
Activities: Over 4 million spectra have been acquired from approx. 460 samples. Data handling and analysis is on-going. The data were acquired using a tissue microarray with no restrictions on age or prior PSA reading.

TASK 2E: DEVELOP PROTOCOLS AND VALIDATE DISTINCTION BETWEEN BENIGN-APPEARING AND ATYPICAL TISSUE

We were able to accomplish task 2e entirely and a manuscript has been submitted (under review). An invention disclosure was filed with the office and technology management, who then decided to file a preliminary paten on the work.

We develop a new fully-automated method to classify cancer versus non-cancer prostate tissue samples. The classification algorithm uses morphological features – geometric properties of epithelial cells/nuclei and lumens – that are quantified based on H&E stained images as well as FT-IR images of the samples. By restricting the features used to geometric measures, we sought to mimic the pattern recognition process employed by human experts, and achieve a robust classification procedure that can produce consistently high accuracy across independent data sets. We systematically evaluate the performance of the new method through cross-validation, and examine its robustness across data sets. We also summarize the specific morphological features that prove to be most informative in classification.

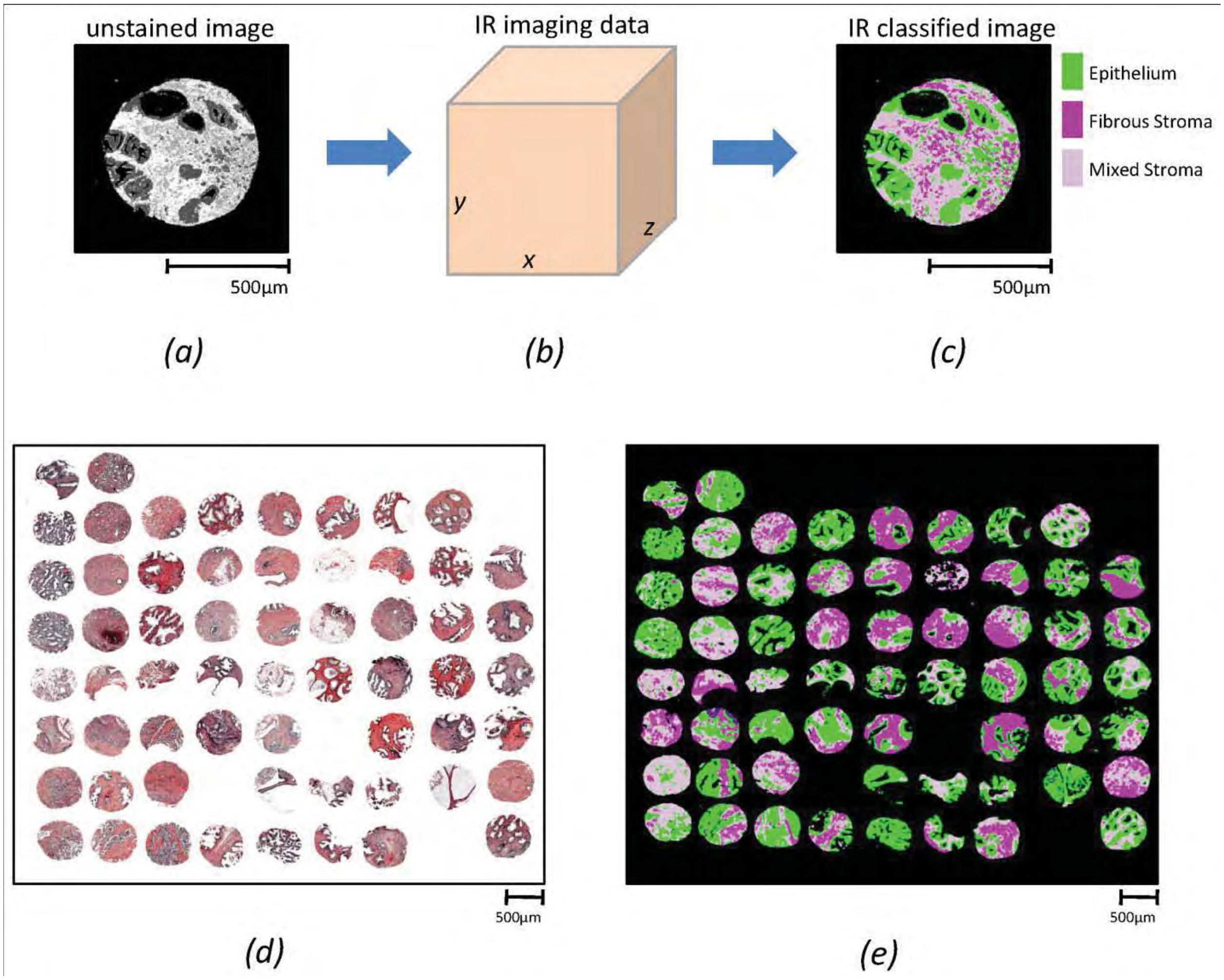


Figure 5. IR imaging data and its use in histologic classification. (Upper row) IR imaging data (b) is acquired for an unstained tissue section (a). The data is then classified into cell types and a classified image (c) is obtained. The colors indicate cell types in a histologic model of prostate tissue. This method is robust and applied to hundreds of tissue samples using the tissue microarray (TMA) format. (Lower row) H&E (d) and IR classified (e) images of a part of the TMAs used.

Methods: Several new methods were developed to accomplish the task.

We begin with a description of the computational pipeline. As noted above, a key aspect of our approach is the use of FT-IR imaging data on a serial section that is H&E-stained to enhance the segmentation of nuclei and lumens. The first two components of the pipeline (§1-2) are geared to this functionality, while the next three components (§3-5) exploit the segmented features obtained from image data to classify the tissue sample (Figure 3).

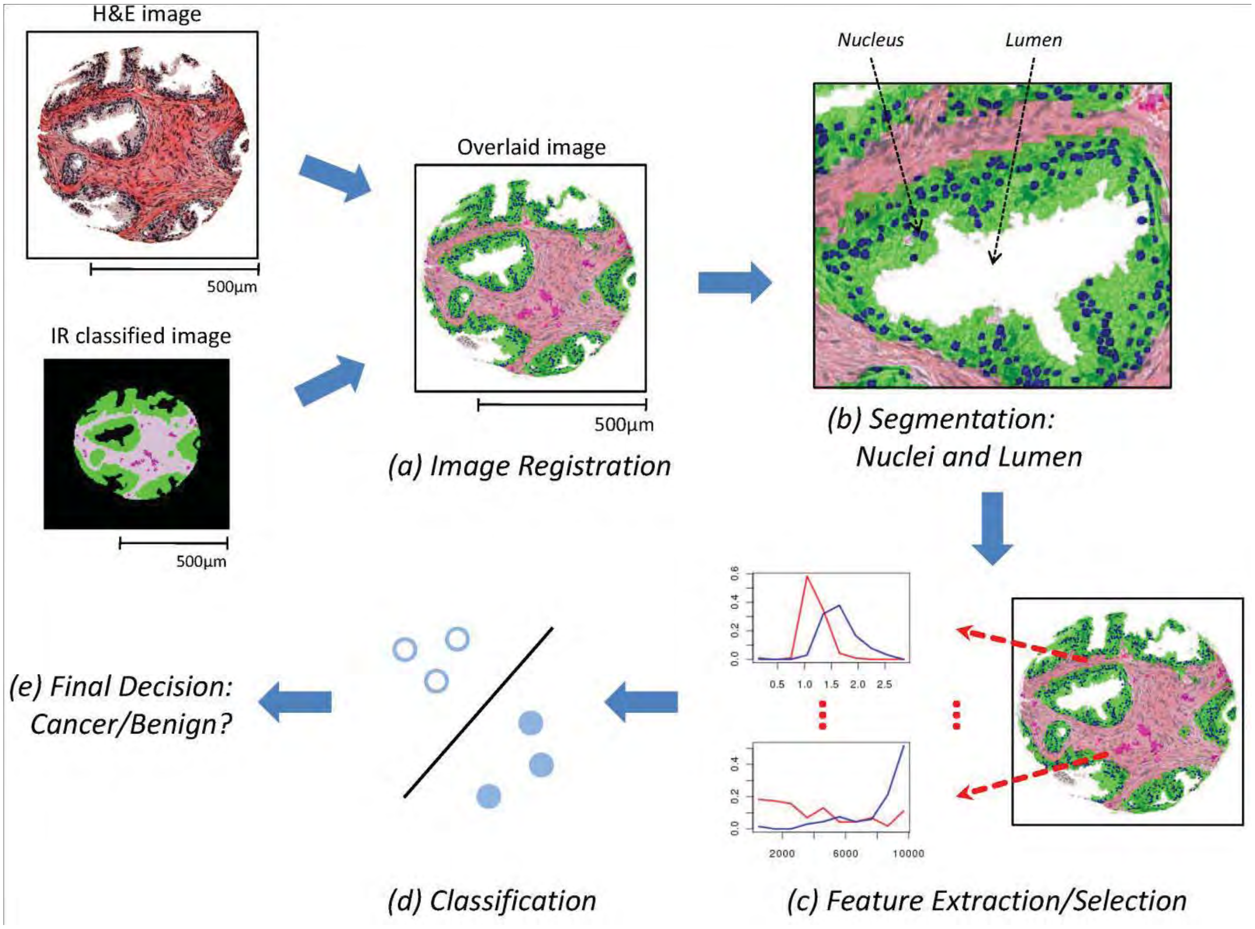


Figure 6. Overview of the approach. (a, b) FTIR spectroscopic imaging data-based cell-type classification (IR classified image), is overlaid with H&E stained image (a), leading to segmentation of nuclei and lumens in a tissue sample (b). (c,d,e) Features are extracted and selected (c), and used by the classifier (d) to predict (e) whether the sample is cancerous or benign.

1. Image Registration

Given two images, the image registration problem can be defined as finding the optimal spatial and intensity transformation of one image to the other. Here, two images are H&E stained and "IR classified" images which were acquired from adjacent tissue samples. The IR classified image represents the FT-IR imaging data, processed as indicated in Figure 2, to classify each pixel as a particular cell type. Although the two samples were physically in the same intact tissue and are structurally similar, the two images have different properties (total image and pixel sizes, contrast mechanisms and data values). Hence, features to spatially register the images are not trivial. The H&E image provides detailed morphological information that could ordinarily be used for registration, but the IR image lacks such information. On the other hand, the IR image specifies the exact areas corresponding to each cell type, but the difficulty in precisely extracting such regions from the H&E image hinders us from using cell-type information for registration. The only obvious features are macroscopic sample shape and empty space (lumens) inside the samples. To utilize these two features and to avoid problems due to differences in the two imaging techniques, both images are first converted into binary images. Due to the binarization, the intensity transformation is not necessary. As a spatial transformation, we use an affine transformation (f) where a coordinate (x_1, y_1) is transformed to the (x_2, y_2) coordinate after translations (t_x, t_y), rotation by θ, and scaling by factor s.

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} t_x \\ t_y \end{bmatrix} + s \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$$

Accordingly, we find the optimal parameters of the affine transformation that minimizes the absolute intensity difference between two images ($I_{reference}$ and I_{target}). In other words, image registration amounts to finding the optimal parameter values $(t_x^*, t_y^*, \theta^*, s^*) = \underset{t_x, t_y, \theta, s}{\arg\min} \left| I_{reference} - f\left(I_{target}; t_x, t_y, \theta, s\right) \right|$. The downhill simplex method is applied to solve the above equation. An example of this registration process is shown in Figure 4.

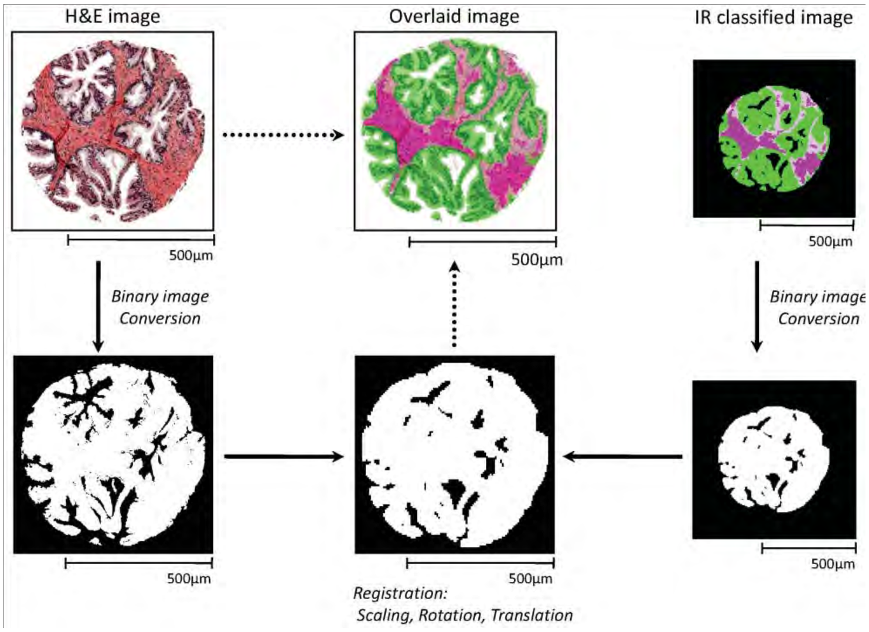


Figure 4. Image Registration. H&E stained images and IR classified images are first converted into binary images. The IR classified image is overlaid with the H&E stained image by affine transformation, with the optimal matching being found by minimizing the absolute intensity difference between two images. After registration, original annotations (color and/or cell-type information) of each image are restored

2. Identification of epithelial cells and their morphologic features

While a number of factors are known to be transformed in cancerous tissues, epithelial morphology is utilized as the clinical gold standard. Hence, we focus here on cellular and nuclear morphology of epithelial nuclei and lumens. These structures are different in normal and cancerous tissues, but are not widely used in automated analysis due to a few reasons. First, as described above, simple detection of epithelium from H&E images is difficult. Second, detection of epithelial nuclei may be confounded by a stromal response that is not uniform for all grades and types of cancers. We focused first on addressing these two challenges that hinder automatically parsing morphologic features such as the size and number of epithelial nuclei and lumens, distance from nuclei to lumens, geometry of the nuclei and lumens, and others (§3). In order to use these properties, the first step is to detect nuclei and lumens correctly and we sought to develop a robust strategy for the same.

2.1. Lumen Detection

In H&E stained images, lumens are recognized to be empty white spaces surrounded by epithelial cells. In normal tissues, lumens are larger in diameter and can have a variety of shapes. In cancerous tissues, lumens are progressively smaller with increasing grade and generally have less distorted elliptical or circular shapes. Our strategy to detect lumens was to find empty areas that are located next to the areas rich in epithelium. White spots inside the sample can be found from the H&E image, and the pixels corresponding to epithelial cells can be mapped on the H&E image from the IR classified image through image registration. We note that while lumens are ideally completely surrounded by epithelial cells (called complete lumens), some samples have lumens (called incomplete lumens) that violate this criterion because only a part of lumen is present in the sample. To identify these incomplete lumens, we use heuristic criteria based on the size, shape, presence of epithelial cells and background around the areas, and distance from the center of the tissue. (See Supplementary Materials for details.)

2.2. Nucleus Detection – single epithelial cells

Epithelial nucleus detection by automated analysis is more difficult than lumen detection due to variability in staining and experimental conditions under which the entire set of H&E images were acquired. Differences between normal and cancerous tissues, and among different grades of cancerous tissues, also hamper facile detection. To handle such variations and make the contrast of the images consistent, we perform smoothing and adaptive histogram equalization prior to nuclei identification. Nuclei are relatively dark and can be modeled as small elliptical areas in the stained images. This geometrical model is often confounded as multiple nuclei can be so close as to appear like one large, arbitrary-shaped nucleus. Also, small folds or edge staining around lumens can make the darker shaded regions difficult to analyze. Here, we exploit the information provided by the IR classified image to limit ourselves to epithelial cells, and use a thresholding heuristic on a color space-transformed image to identify nuclei with high accuracy. Epithelial pixels that are identified on the H&E images using the IR overlay provide pixels of dominated by one of two colors: blue or pink, which arise from the nuclear and cytoplasmic component respectively. For nuclei restricted to epithelial cells in this manner, a set of general observations were made that led us to convert the stained image to a new color space "RG–B" ($|R + G - B|$). (R, G, and B represent the intensity of Red, Green, and Blue channels, respectively.) This transformation, followed by suitable thresholding, was able to successfully characterize the areas where nuclei are present. The threshold values are adaptively determined for Red and Green channels due to the variations in the color intensity. (See Supplementary Materials for details.) Finally, filling holes and gaps within nuclei by a morphological closing operation, the segmentation of each nucleus is accomplished by using a watershed algorithm followed by elimination of false detections. The size, shape, and average intensity are considered to identify and remove artifactual nuclei. Figure 5 details the nucleus detection procedure.

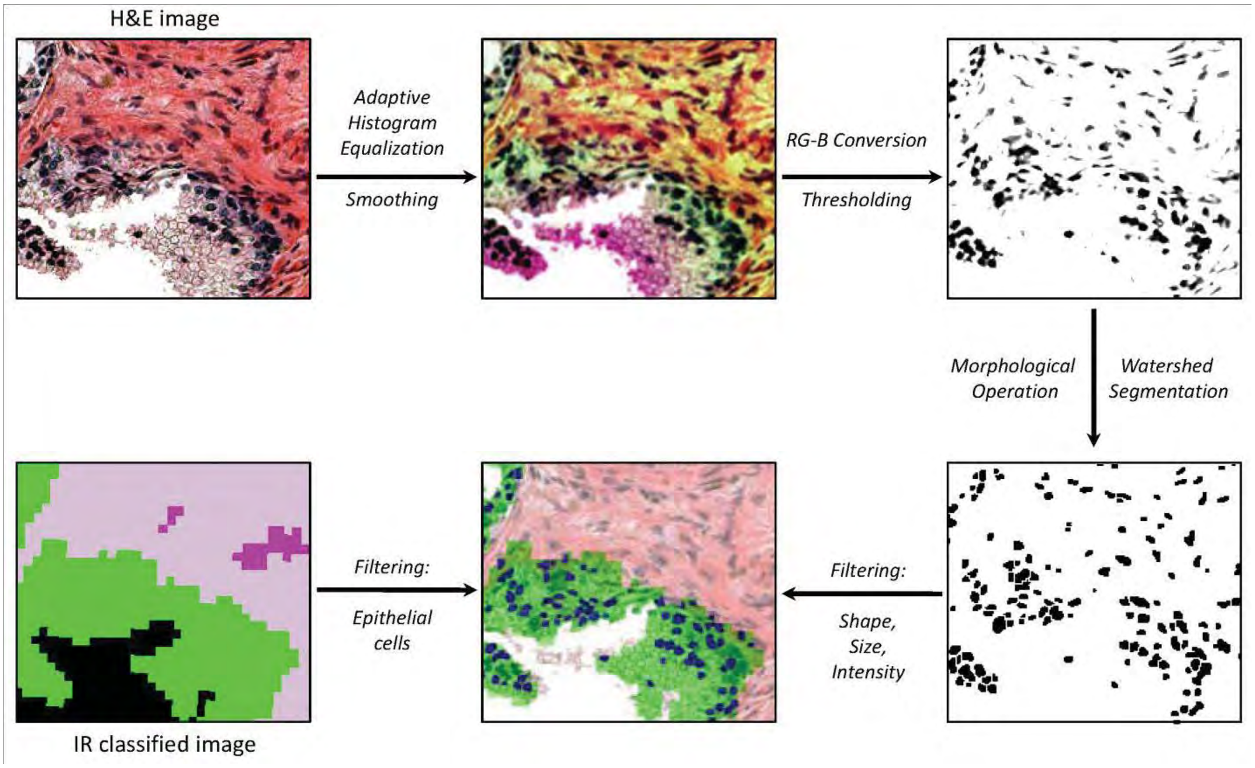


Figure 7. Nucleus Detection. Smoothing and adaptive histogram equalization are performed to alleviate variability in H&E stained image and to obtain better contrast. "RG – B" conversion followed by thresholding characterizes the areas where nuclei exist. Morphological closing operation is performed to fill holes and gaps within nuclei, and a watershed algorithm segments each individual nuclei. The segmented nuclei are constrained by their shape, size, and average intensity and epithelial cell classification (green pixels) provided by the overlaid IR image.

3. Feature Extraction

As mentioned above, the characteristics of nuclei and lumens change in cancerous tissues. In a normal tissue, epithelial cells are located mostly in thin layers around lumens. In cancerous tissue, these cells generally grow to fill lumens, resulting in a decrease in the size of lumens, with the shape of lumens becoming more elliptical or circular. The epithelial association with a lumen becomes inconsistent and epithelial foci may adjoin lumens or may also exist without an apparent lumen. Epithelial cells invading the extra-cellular matrix also result in a deviation from the well-formed lumen structure; this is well-recognized as a hallmark of cancer. Due to filling lumen space and invasion into the extra-cellular space, the number density of epithelial cells increases in tissue. The size of individual epithelial cells and their nuclei also tend to increase as malignancy of a tumor increases. Motivated by such recognized morphological differences between normal and cancerous tissues, we chose to use epithelial nuclei and lumens as the basis of the several quantitative features that our classification system works with. (See examples of such features in Figure 6.) It is notable that these observations are qualitative in actual clinical practice and have not been previously quantified.

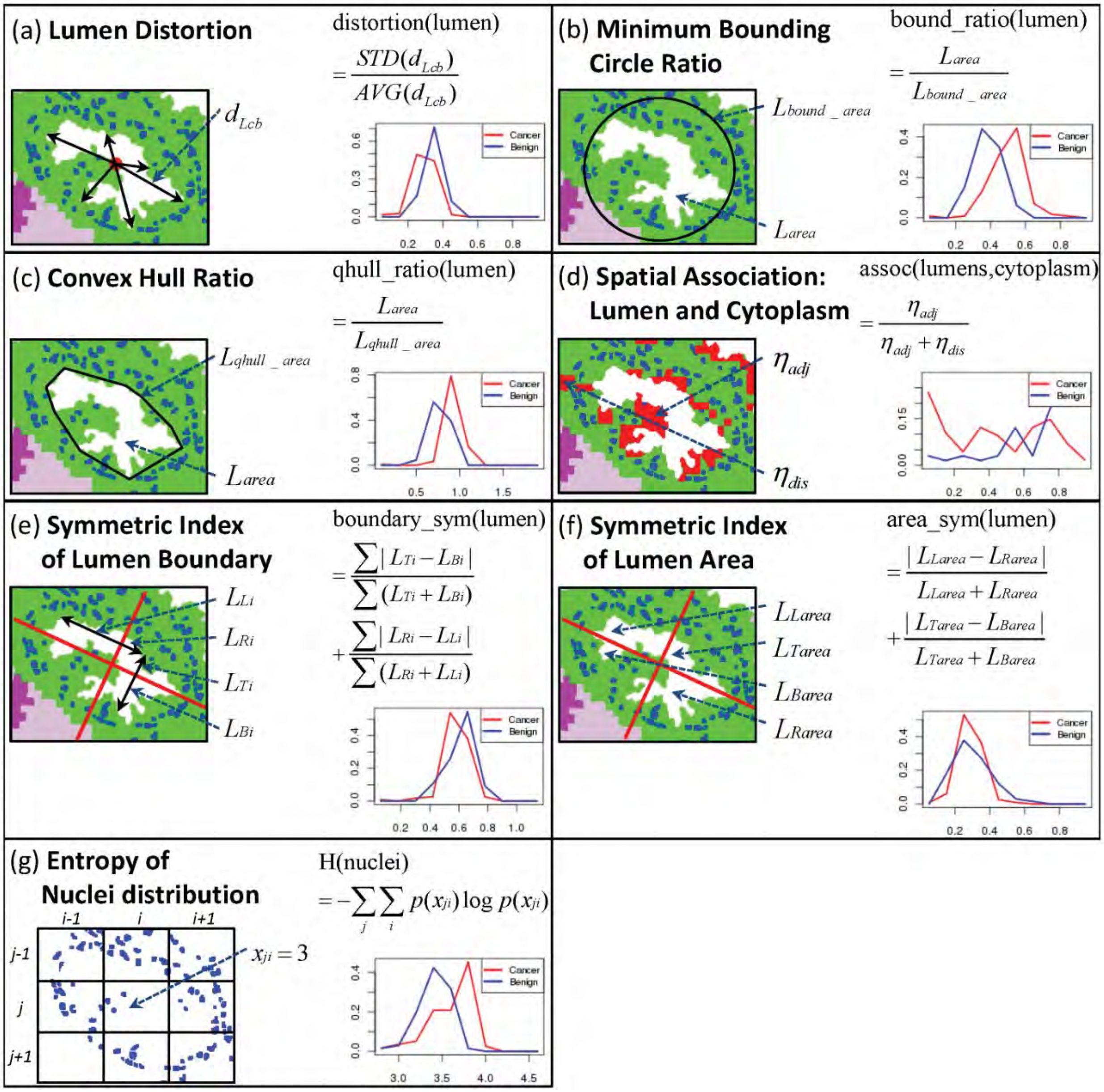


Figure 8. Examples Features. Each panel shows one example feature, along with the distributions of the feature's values for cancer (red) and benign (blue) classes.

3.1. Epithelial cell-related features

We use epithelial cell type classification from IR data to measure epithelium-related features. However, individual epithelial cells in the tissue are not easily delineated. Therefore, in addition to features directly describing epithelial cells, we also quantify properties of epithelial nuclei, which are available from the segmentation described in §2. The quantities we measure in defining features are: (1) size of epithelial cells, (2) size of epithelial nuclei, (3) number of nuclei in the sample, (4) distance from a nucleus to the closest lumen, (5) distance from a nucleus to the epithelial cell boundary, (6) number of "isolated" nuclei (nuclei that have no neighboring nucleus within a certain distance), (7) number of nuclei located "far" from lumens, and (8) entropy of spatial distribution of nuclei (Figure 6G). Supplementary Materials provide specifics of these measures and their calculation.

3.2. Lumen-related features

Features describing glands have been shown to be effective in PCa classification. Here, we try to characterize lumens and mostly focus on the differences in the shape of the lumens. The quantities we measure in defining these features are: (1) size of a lumen, (2) number of lumens, (3) lumen "roundness", defined as $\frac{L_{peri}}{2L_{area}}r$ where L_{peri} is the perimeter of the lumen, L_{area} is the size of the lumen, and *r* is the radius of a circle of size L_{area}, (4) lumen "distortion" (Figure 6A), computed as $\frac{STD(d_{L_{cb}})}{AVG(d_{L_{cb}})}$ where $d_{L_{cb}}$ is the distance from the center of a lumen to the boundary of the lumen and *AVG(·)* and *STD(·)* represent the average and standard deviation, (5) lumen "minimum bounding circle ratio" (Figure 6B), defined as the ratio of the size of a minimum bounding circle of a lumen to the size of the lumen, (6) lumen "convex hull ratio" (Figure 6C), which is the ratio of the size of a convex hull of a lumen to the size of the lumen, (7) symmetric index of lumen boundary (Figure 6E, see Supplementary Materials), (8) symmetric index of lumen area (Figure 6F, see Supplementary Materials), and (9) spatial association of lumens and cytoplasm-rich regions (Figure 6D, see Supplementary Materials). Features (3) – (8) are various ways to summarize lumen shapes, while feature (9) is motivated by the loss of functional polarization of epithelial cells in cancerous tissues.

3.3. Global & local tissue features

We have described above the individual measures of epithelium and lumen related quantities that form the basis of the features used by our classification system. Normally, these features have to be summary measures over the entire tissue sample or desired classification area. Hence, we employ average (AVG) or standard deviation (STD), and in some cases the sum total (TOT) of these quantities for further analysis. These features are called "global" features since they are calculated from the entire sample. However, in some cases global features may be misleading, especially where only a part of the tissue sample is indicative of cancer. Therefore, in addition to global features, we define "local" features by sliding a rectangular window of a fixed size (typically 100x100 pixels) throughout a tissue sample, computing the average or sum total of the feature in each window, and computing the standard deviation and/or extrema over the values for all windows (Figure 7). In all, 67 features (29 global and 38 local features) are defined capturing various aspects of tissue morphology**.**

4. Feature Selection

Feature selection is the step where the classifier examines all available features (67 in our case) with respect to the training samples, and selects a subset to use on test data. This selection is generally based on the criterion of high accuracy on training data, but also strives to ensure generalizability beyond the training data. We adopt a two-stage feature selection approach here. In the first stage, we generate a set of candidate features ($C_{candidate}$) by using the so-called minimum-redundancy-maximal-relevance (mRMR) criterion. In each iteration, given a feature set chosen thus far, mRMR chooses the single additional feature that is least redundant with the chosen features, while being highly correlated with the class label. $C_{candidate}$ is a set of features that is expected to be close to the optimal feature set for a dataset and a classifier under consideration. It is constructed as follows. Given a feature set F = (f_1, …, f_M) ordered by mRMR, AUC of the set of *i* top-ranked features is computed for varying values of *i*. We limit the value of *i* to be ≤ 30. The feature subset with the best AUC is chosen as the $C_{candidate}$. In the second stage, feature selection continues with $C_{candidate}$ as the starting point, using the sequential floating forward selection (SFFS) method. This method sequentially adds new features followed by conditional deletion(s) of already selected features. Starting with the $C_{candidate}$, SFFS searches for a feature $x \notin C_{candidate}$ that maximizes the AUC among all feature sets $C_{candidate} \cup \{x\}$, and adds it to $C_{candidate}$. Then, it finds a feature $x \in C_{candidate}$ that maximizes the AUC among all feature sets $C_{candidate} - \{x\}$. If the removal of x improves the highest AUC obtained by $C_{candidate}$, x is deleted from $C_{candidate}$. As long as this removal improves upon the highest AUC obtained so far, the removal step is repeated. SFFS repeats the addition and removal steps until AUC reaches 1.0 or the number of additions and deletions exceeds 20, and the feature set with the highest AUC thus far is chosen as the optimal feature set. The classification capability of a feature set, required for feature selection, is measured by the area under the ROC curve (AUC), obtained by cross-validation on the training set.

5. Classification

We note that there are two levels of classification here. In the first, IR spectral data is used to provide histologic images where each pixel has been classified as a cell type. In the second, the measures from H&E images and IR images are used to classify tissue into disease states. In this manuscript, we do not discuss the first classification task as its development and results are well-documented. For the latter task, we used a well established classification algorithm, namely support vector machine (SVM). Two cost factors are introduced to deal with an imbalance in training data. The ratio between two cost functions was chosen as

$$\frac{C_+}{C_-} = \frac{\text{number of negative training examples}}{\text{number of positive training examples}}$$

to make the potential total cost of the false positives and the false negatives the same. (See Supplementary Materials for details.)

6. Data preparation

All of the H&E stained images were acquired on a standard optical microscope at 40x magnification. The size of each pixel is 0.9636 um x 0.9636 um. On the other hand, the pixel size of IR images is 6.25um x 6.25um. The acquisition was previously described in previous years' reports. Two data sets, stained under different conditions, were used in this study. The first dataset ("Data1") consists of 66 benign samples and 115 cancer samples, and the second set ("Data2") includes 14 benign and 36 cancer samples. These were previously acquired under the grant.

Results and discussion: We then applied the methods to classify prostate tissue and the results are presented below.

1. The classification system achieves AUC greater than 0.97 on both data sets

We first performed *K*-fold cross validation on each dataset. The data set was divided into *K* roughly equal-sized partitions, one partition was left out as the "test data", the classifier was trained on the union of the remaining *K – 1* partitions (the "training data") and evaluated on the test data. This was repeated *K* times, with different choices of the left-out partition. (We set *K* = 10.) In each repetition, cross-validation on the training data was used to select the feature set with the highest AUC as explained in §4. The correct and incorrect predictions in the test data, across all *K* repetitions, were summarized into a ROC plot and the AUC was computed, along with specificities when sensitivity equals 90, 95, or 99%. Since the cross-validation exercise makes random choices in partitioning the data set, we examined averages of these performance metrics over 10 repeats of the entire cross validation pipeline. The average AUC for *Data1* and *Data2* were 0.982 and 0.974 respectively (Table 1, "feature extraction" = "IR & HE"). At 90%, 95%, and 99% sensitivities, the average specificity achieved on *Data1* was 94.76%, 90.91%, and 77.80% respectively, while that on *Data2* was 92.53%, 84.19%, and 49.54% respectively.

One way to interpret the above values is to examine our automated pipeline as a pre-screening mechanism to identify the samples to be examined by a human pathologist. At a "true positive rate" of 99% (which means that only 1% of the cancer samples will be missed by the screen), the "false positive rate" is 22.2% (i.e., 22.2% of the benign samples will make it through the screen) on average for *Data1* (Table1), thereby reducing the workload of the pathologist by 4.5-fold. While the error rate of manual pathology determinations is generally accepted to be in 1-5% range, inclusion of confounding cancer mimickers raises the rate to as high as 7.5%. Also noteworthy is the observation that the same algorithm performs consistently well on both data sets, that were obtained from different staining conditions. This speaks to the robustness of the classification framework, an attribute that we investigated further in the next exercise.

2. Classification system is robust to staining conditions

Here, we trained a classifier on *Data1* and tested its performance on *Data2*. We observed an average AUC of 0.956, with average specificity of 88.57%, 81.92%, and 26.86% at sensitivity equaling 90%, 95%, and 99% respectively (Table 2, "feature extraction" = "IR & HE"). These values are competitive with the cross-validation results on *Data2* (Table 1), where the training and testing were both performed on (disjoint parts of) *Data2*.

3. IR data is critical to classification performance

To assess the utility of the IR-based cell-type classification, we repeated the above exercises after extracting features without the guidance of the IR data; i.e., epithelial cells were predicted from the H&E images alone (see Supplementary Materials for details). All of the features defined in §3 were used, except for "Spatial association of lumens and apical regions", since the distinction between cytoplasm-rich and nuclear-rich region in epithelial cells was unclear in H&E images. The results from this disadvantaged classifier are shown in Tables 1 and 2 ("feature extraction" = "HE only"). For both types of experiments, we obtained lower average AUCs and specificity values. For instance, the AUC of cross-validation in *Data2* (Table 1) dropped from 0.974 to 0.880. Similarly, the results of validation between datasets (Table 2) were substantially worse now compared to the IR-guided classification, with the AUC dropping from 0.956 to 0.918. This indicates that feature extraction with the help of the IR cell-type classification is critical to consistent and reliable classification of cancer versus benign tissue samples.

Dataset	Feature Extraction	AUC		Sensitivity (%)	Specificity (%)		M_f
		AVG	STD		AVG	STD	

Data1	IR & HE	0.982	0.0030	90	94.76	1.64	13
				95	90.91	1.62	
				99	77.80	5.52	
	HE only	0.968	0.0052	90	91.64	2.26	11
				95	83.90	1.91	
				99	53.43	13.65	
Data2	IR & HE	0.974	0.0145	90	92.53	7.11	7
				95	84.19	10.84	
				99	49.54	22.51	
	HE only	0.880	0.0175	90	61.34	10.31	8
				95	22.21	10.06	
				99	11.21	6.01	

Table 1 . Classification results via cross-validation.
AVG and STD denote average and standard deviation across ten repeats of cross-valdiation. M_f is the median size of the feature set obtained by feature selection from training data. Column "Feature Extraction" indicates if features were obtained using H&E as well as IR data, or with H&E data alone.

Feature Extraction	Dataset	AUC		Sensitivity (%)	Specificity (%)		M_f
		AVG	STD		AVG	STD	
IR & HE	Train	0.994	0.0006	90	98.30	0.68	13
				95	96.58	1.10	
				99	91.55	2.55	
	Test	0.956	0.0089	90	88.57	5.96	
				95	81.92	5.28	
				99	26.86	15.50	
HE only	Train	0.986	0.0021	90	97.77	0.97	10
				95	91.56	2.49	
				99	79.29	4.47	
	Test	0.918	0.0100	90	65.51	8.37	
				95	46.14	7.53	
				99	13.29	6.94	

Table 2. Validation between datasets.
A classifier is trained on *Data1* and tested on *Data2*. AVG and STD denote the average and standard deviation. M_f is the median size of the optimal feature set. Column "Feature Extraction" indicates if features were obtained using H&E as well as IR data, or with H&E data alone. Column "Dataset" indicates if the performance metrics are from training data (*Data1*) or from test data (*Data2*).

Previously, Tabeshi *et al*. achieved an accuracy of 96.7% via cross validation in cancer/no-cancer classification. Color, morphometric, and texture features were extracted, and all images were acquired under similar conditions. We note that our classification result (Table 1), based solely on morphology, is comparable to their result; however the software developed by Tabeshi et al. was not available for evaluation in our data sets. Color and texture features could provide additional information; however, their robustness to different data sets is questionable, and their interpretation is not as obvious as that of morphological features, which are used in clinical practice. Different data sets may have varied properties which may be attributable to staining variations, inconsistent image acquisition settings, and image preparation. The performance of the same method based on texture features has been seen to greatly change from one data set to another. Variations in staining may affect color features. In contrast, morphological features were shown to be robust to varying image acquisition settings. Nonetheless, the quality of morphological features is subject to segmentation of histologic objects. Thus, any method based on morphological features will benefit from the IR cell-type classification.

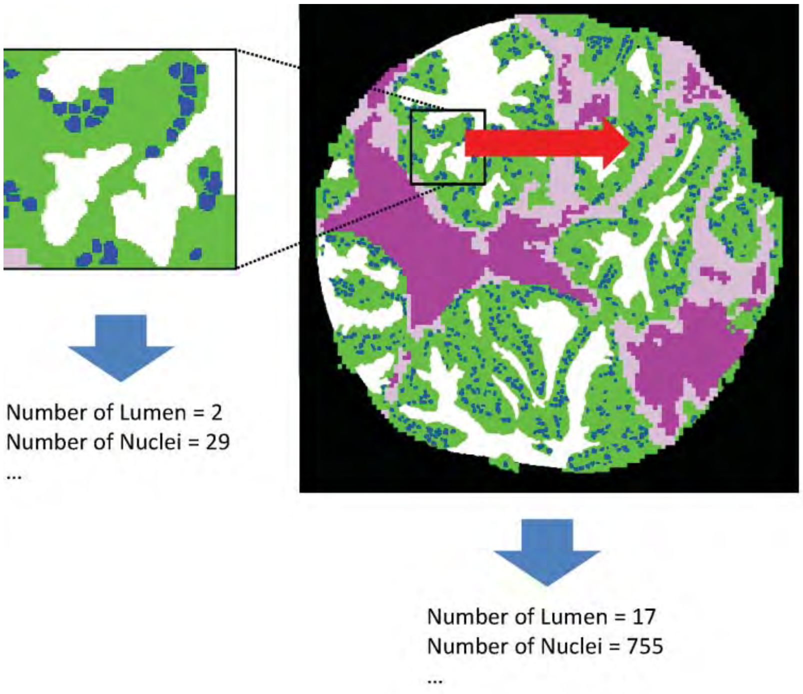


Figure 9. Global and Local Feature Extraction. Global features are extracted from the entire tissue sample, and local features are extracted by sliding a window of a fixed size across the tissue sample and computing summary statistics, such as standard deviation, of window-specific scores. In this example, the global feature "number of nuclei" has value 755, while one example position of the sliding window is shown, with "number of nuclei" = 29.

4. Examination of discriminative features

We examined the importance of each feature by its rank in the first phase of feature selection, based on its "relevance" to the class label (see Supplementary Materials, mRMR). Since different features (e.g., average or standard deviation, global or local features) based on the same underlying quantity (e.g., "lumen roundness") generally have similar relevance, we examined the average relevance of features in each of 17 feature categories (Figure 8), for each data set. The complete list of the individual features and their relevance and mRMR rank (for *Data1*) is available in Figure 9. For *Data1*, lumen-related feature categories are most relevant in general, while epithelium-related feature categories are most important for *Data2*. It is surprising that the top 3 feature categories in *Data1* (Figure 8, blue bars) – size of lumen, lumen roundness, and lumen convex hull ratio – have very low relevance in *Data2*, although we note that this may be in large part due to variations in staining and malignancy of tumors between the two data sets. Also, examining the features (or feature categories) with highest relevance alone may be slightly misleading, because this examination does not account for redundancy among features.

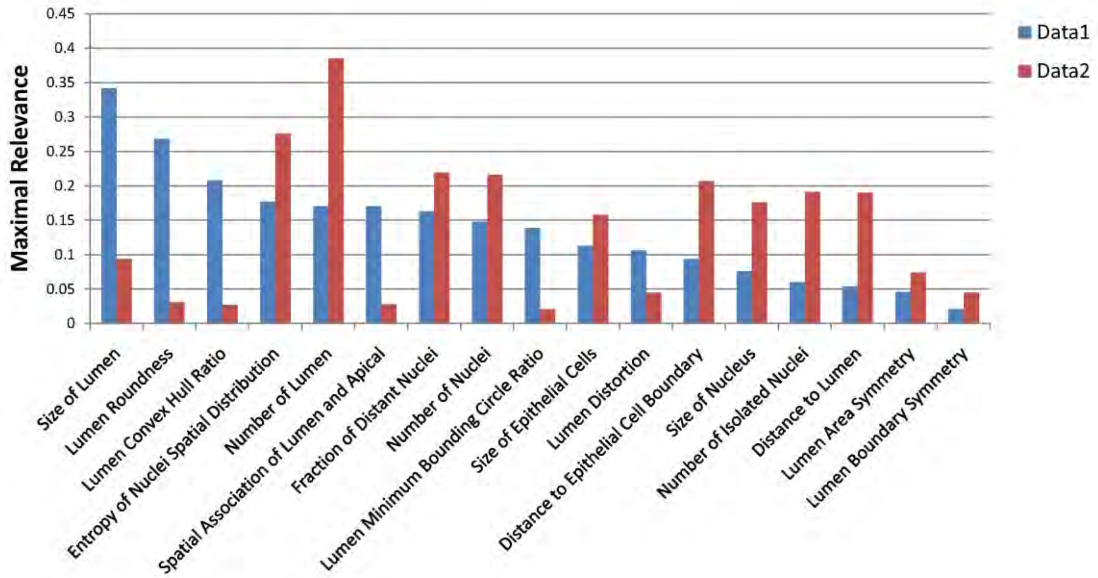
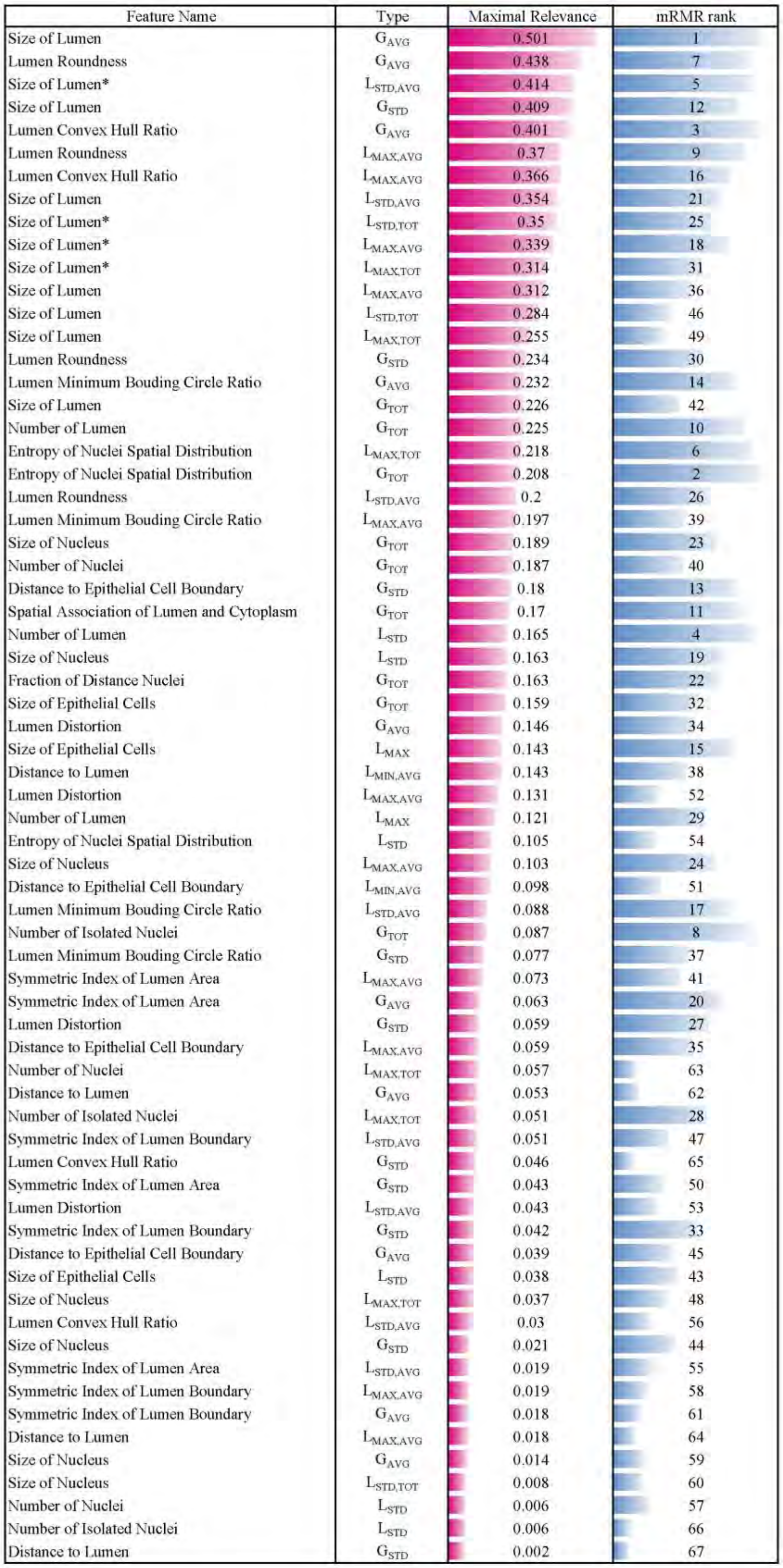


Figure 10. Importance of 17 feature categories. The average "maximal relevance" of features belonging to each feature category is shown, for both data sets, sorted in decreasing order for the first data set.



Feature Name	Type	Maximal Relevance	mRMR rank
Size of Lumen	G_{AVG}	0.501	1
Lumen Roundness	G_{AVG}	0.438	7
Size of Lumen*	$L_{STD,AVG}$	0.414	5
Size of Lumen	G_{STD}	0.409	12
Lumen Convex Hull Ratio	G_{AVG}	0.401	3
Lumen Roundness	$L_{MAX,AVG}$	0.37	9
Lumen Convex Hull Ratio	$L_{MAX,AVG}$	0.366	16
Size of Lumen	$L_{STD,AVG}$	0.354	21
Size of Lumen*	$L_{STD,TOT}$	0.35	25
Size of Lumen*	$L_{MAX,AVG}$	0.339	18
Size of Lumen*	$L_{MAX,TOT}$	0.314	31
Size of Lumen	$L_{MAX,AVG}$	0.312	36
Size of Lumen	$L_{STD,TOT}$	0.284	46
Size of Lumen	$L_{MAX,TOT}$	0.255	49
Lumen Roundness	G_{STD}	0.234	30
Lumen Minimum Bouding Circle Ratio	G_{AVG}	0.232	14
Size of Lumen	G_{TOT}	0.226	42
Number of Lumen	G_{TOT}	0.225	10
Entropy of Nuclei Spatial Distribution	$L_{MAX,TOT}$	0.218	6
Entropy of Nuclei Spatial Distribution	G_{TOT}	0.208	2
Lumen Roundness	$L_{STD,AVG}$	0.2	26
Lumen Minimum Bouding Circle Ratio	$L_{MAX,AVG}$	0.197	39
Size of Nucleus	G_{TOT}	0.189	23
Number of Nuclei	G_{TOT}	0.187	40
Distance to Epithelial Cell Boundary	G_{STD}	0.18	13
Spatial Association of Lumen and Cytoplasm	G_{TOT}	0.17	11
Number of Lumen	L_{STD}	0.165	4
Size of Nucleus	L_{STD}	0.163	19
Fraction of Distance Nuclei	G_{TOT}	0.163	22
Size of Epithelial Cells	G_{TOT}	0.159	32
Lumen Distortion	G_{AVG}	0.146	34
Size of Epithelial Cells	L_{MAX}	0.143	15
Distance to Lumen	$L_{MIN,AVG}$	0.143	38
Lumen Distortion	$L_{MAX,AVG}$	0.131	52
Number of Lumen	L_{MAX}	0.121	29
Entropy of Nuclei Spatial Distribution	L_{STD}	0.105	54
Size of Nucleus	$L_{MAX,AVG}$	0.103	24
Distance to Epithelial Cell Boundary	$L_{MIN,AVG}$	0.098	51
Lumen Minimum Bouding Circle Ratio	$L_{STD,AVG}$	0.088	17
Number of Isolated Nuclei	G_{TOT}	0.087	8
Lumen Minimum Bouding Circle Ratio	G_{STD}	0.077	37
Symmetric Index of Lumen Area	$L_{MAX,AVG}$	0.073	41
Symmetric Index of Lumen Area	G_{AVG}	0.063	20
Lumen Distortion	G_{STD}	0.059	27
Distance to Epithelial Cell Boundary	$L_{MAX,AVG}$	0.059	35
Number of Nuclei	$L_{MAX,TOT}$	0.057	63
Distance to Lumen	G_{AVG}	0.053	62
Number of Isolated Nuclei	$L_{MAX,TOT}$	0.051	28
Symmetric Index of Lumen Boundary	$L_{STD,AVG}$	0.051	47
Lumen Convex Hull Ratio	G_{STD}	0.046	65
Symmetric Index of Lumen Area	G_{STD}	0.043	50
Lumen Distortion	$L_{STD,AVG}$	0.043	53
Symmetric Index of Lumen Boundary	G_{STD}	0.042	33
Distance to Epithelial Cell Boundary	G_{AVG}	0.039	45
Size of Epithelial Cells	L_{STD}	0.038	43
Size of Nucleus	$L_{MAX,TOT}$	0.037	48
Lumen Convex Hull Ratio	$L_{STD,AVG}$	0.03	56
Size of Nucleus	G_{STD}	0.021	44
Symmetric Index of Lumen Area	$L_{STD,AVG}$	0.019	55
Symmetric Index of Lumen Boundary	$L_{MAX,AVG}$	0.019	58
Symmetric Index of Lumen Boundary	G_{AVG}	0.018	61
Distance to Lumen	$L_{MAX,AVG}$	0.018	64
Size of Nucleus	G_{AVG}	0.014	59
Size of Nucleus	$L_{STD,TOT}$	0.008	60
Number of Nuclei	L_{STD}	0.006	57
Number of Isolated Nuclei	L_{STD}	0.006	66
Distance to Lumen	G_{STD}	0.002	67

Figure 11. List of features and their maximal relevance and “mRMR rank”. In the second column, *G* and *L* represent global and local features, respectively. *AVG*, *STD*, *TOT*, and

***MAX* denote the average, standard deviation, total amount, and extremal value of features. * In computing local features representing "size of lumen", two options are available: one is to consider only the part of the lumen within the window, and the other is to consider the entire lumen into account. Asterisk indicates that the former option was chosen.**

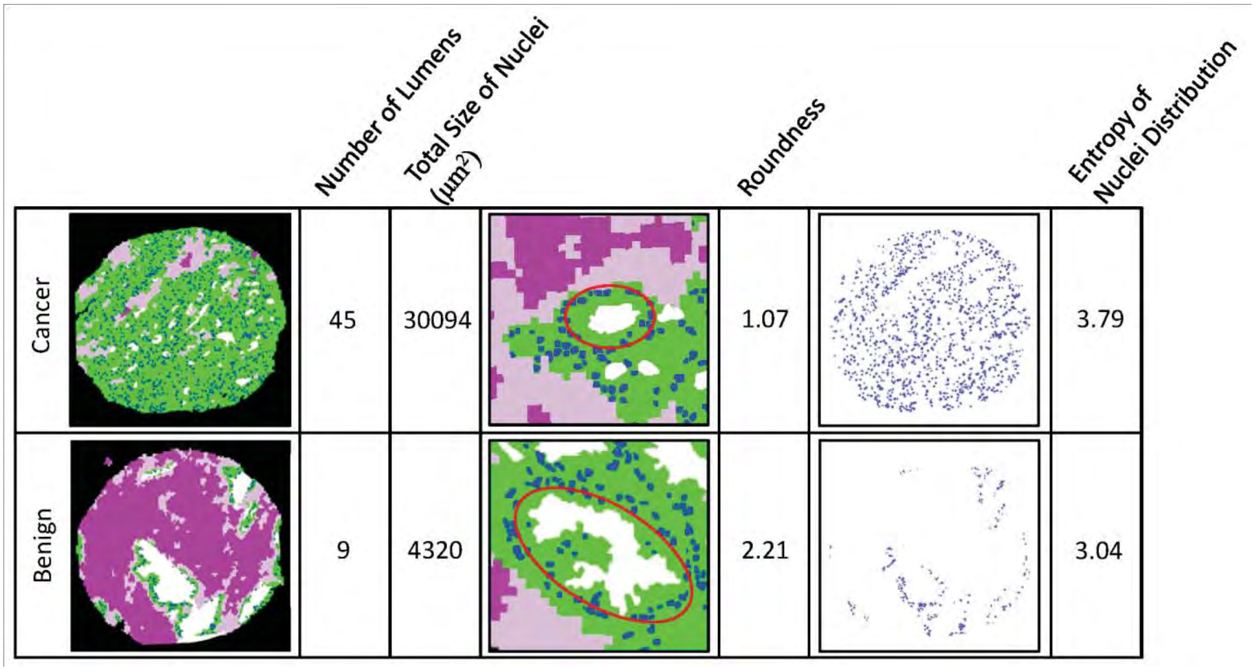


Figure 12. Optimal features for distinguishing cancer and benign tissue samples. The four features shown here are always present in the optimal feature set chosen by the classifier.

Conclusions

In completing this task, we have presented a means to eliminate epithelium recognition deficiencies in classifying H&E images for presence or absence of cancer. The method is entirely transparent to a user and does not involve any adjustment or decision-making based on spectral data. We were able to achieve very effective fusion of the information from two different modalities, namely optical and IR microscopy, that provide very different types of data with different characteristics. Several features of the tissue were quantified and employed for classification. We found that robust classification could be achieved using a few measures, which are detailed to arise from epithelial/lumen organization and provide a reasonable explanation for the accuracy of the model. The choice of combining the IR and optical data is shown to be necessary for achieving the high accuracy values observed. We anticipate that the combined use of the two microscopies – structural and chemical – will lead to an accurate, robust and automated method for determining cancer within biopsy specimens.

TASK 2F: DEVELOP CALIBRATION FOR PREDICTING CANCER GRADE (MONTHS 18-22)

Motivation:

Quality assurance in clinical pathology plays a critical role in the management of patients with prostate cancer as pathology is the gold standard of diagnosis and forms a cornerstone of patient therapy. Methods to integrate quality development, quality maintenance, and quality improvement to ensure accurate and consistent test results are, hence, critical to cancer management in any setting. These factors have a direct bearing on patient outcomes, financial aspects of disease management as well as malpractice concerns. One of the major failings in prostate pathology today is the rate of missed tumors and variability in grading. It is well known that the grading of prostate tissues suffers from intra- and inter-pathologist variability. In the studies of intra- and inter-pathologist reproducibility, the exact intra-pathologist agreement was achieved in 43-78% of the instances, and in 36-81% of the instances, the exact inter-pathologist agreement was reported. It is also known that the variability of the grading could be reduced after pathologists are re-trained. There could be many ways to educate pathologists such as meetings, courses, online tutorials, and etc, but these are not time- and cost-effective for routine everyday decisions. Therefore, building an automated, fast, and objective method to aid pathologists to examine prostate tissues will greatly help to attain reliable and consistent diagnoses. This will reduce healthcare costs and the chances of malpractice lawsuits as well as improve patient outcomes in therapy.

Innovation in our approach and potential benefits:

When a pathologist examines tissue, he/she looks at a stained imaged of tissue and mentally compares it against a database of previous knowledge or information in books. In essence, the pathologist is manually matching structural patterns he/she has seen earlier and mentally recalling the diagnosis made such that he/she can make the same diagnosis in the specific test case. Here, we report developing a computer information and management and decision-making system that relies of one or more measures of the structure of tissue to provide images from a database that are similar to the sample under consideration. We emphasize that the system does not provide a diagnosis but simply provides the closest matching cases that enable a pathologist to make a diagnosis. We also propose here the new idea of constructing a database of pre-examined prostate tissues and providing similar tissue samples with pathologists from the database while they examine an unknown tissue sample. To our knowledge, no such system currently exists. Further, we propose that our system may or may not use infrared chemical imaging data in comparisons. Comparing with the pre-examined tissues samples, we expect that pathologists to make more consistent and accurate decision. As we build a database of prostate tissue samples, we represent each tissue sample by its morphology. Given an unknown tissue sample, the similarities between the unknown sample and the tissue samples in the database are measured based on the morphological properties, and the most similar tissue samples are retrieved. The pathologist may indicate that certain matches were better than others, resulting in an updating of the database and matching algorithms as needed. The updating may be conducted in real-time.

Work accomplished:

Morphological features have been shown to be able to characterize prostate tissues and can be used for the diagnostic purpose. Here, 67 morphological features, which are based on lumens and epithelial nuclei, were extracted from each tissue sample. The database stores the morphological features for the tissue samples which have already been examined by pathologists.

Once we have an unknown prostate tissue sample (query), first of all, the morphological features are extracted from the tissue sample. Secondly, the similarities between the query and the tissue samples in the database are computed using Euclidean distance based on the morphological features. Lastly, the most similar *k* tissue samples to the query are retrieved from the database.

To assess the goodness of the method, we have tested our method on a dataset composed of 181 tissue samples. In the dataset, 5, 23, 66, and 21 tissue samples are Gleason grade 2, 3, 4, and 5 cancer ("*Cancer"*), respectively, and 20 and 46 tissue samples are BPH and normal ("*Benign"*), respectively. Due to the small number of tissue samples, Gleason grade 2 is ignored for the further consideration. As mentioned above, each of tissue samples is represented by 67 morphological features.

In order to measure the performance of the method, we adopted *k*-nearest neighbor (kNN) algorithm and predicted the grade of the query by majority voting. Both accuracy and kappa-coefficient were computed for the predictions. Since pathologists may be more interested in grading of cancerous tissue samples, we also applied our method only to the *"Cancer"* tissue samples; i.e., Gleason grade 3, 4, and 5 samples.

We performed Leave-one-out (LOO) cross-validation on the dataset. LOO leaves one example as a validation data and uses the remaining examples as training data. In our method, the validation data is the query, and the training data is regarded as the database. It should be noted that the number of tissue samples in each grade in the dataset varies. The imbalance in the dataset could affect the prediction made by kNN algorithm. To tackle the problem, we randomly selected the same number of tissue samples from each grade and performed LOO on the sub-dataset. This repeated 100 times, and the average accuracy and kappa-coefficient were computed over the repeats.

Our method is subject to the choice of the number of nearest neighbors to consider for the prediction and the number of features to use for the similarity computation. To examine the effect of them, we computed the average accuracy and kappa-coefficient over 100 repeats as increasing the two factors (Fig. 1). The accuracy decreases as increasing the number of nearest neighbors, and the more features we use, the higher accuracy achieved. The highest average accuracy achieved for grading both "*Cancer"* and "*Benign"* samples (i.e, 5 grades) was 42% using 7 features and 1 nearest neighbor (Fig. 1a). By using 8 features and 1 nearest neighbor, the highest accuracy of 52% achieved for grading only "*Cancer"* samples (i.e., 3 grades) (Fig. 1c). Both cases also achieved the average Kappa coefficient of 0.27 (Fig. 1b, d). In Fig. 2, the distribution of the grade of the retrieved samples is shown. Distinction between *"Cancer"* and *"Benign"* samples is obvious (Fig 2a), but among *"Cancer"*, the retrieved samples often do not belong to the same grade with the query, especially between Gleason grade 3 and 4.

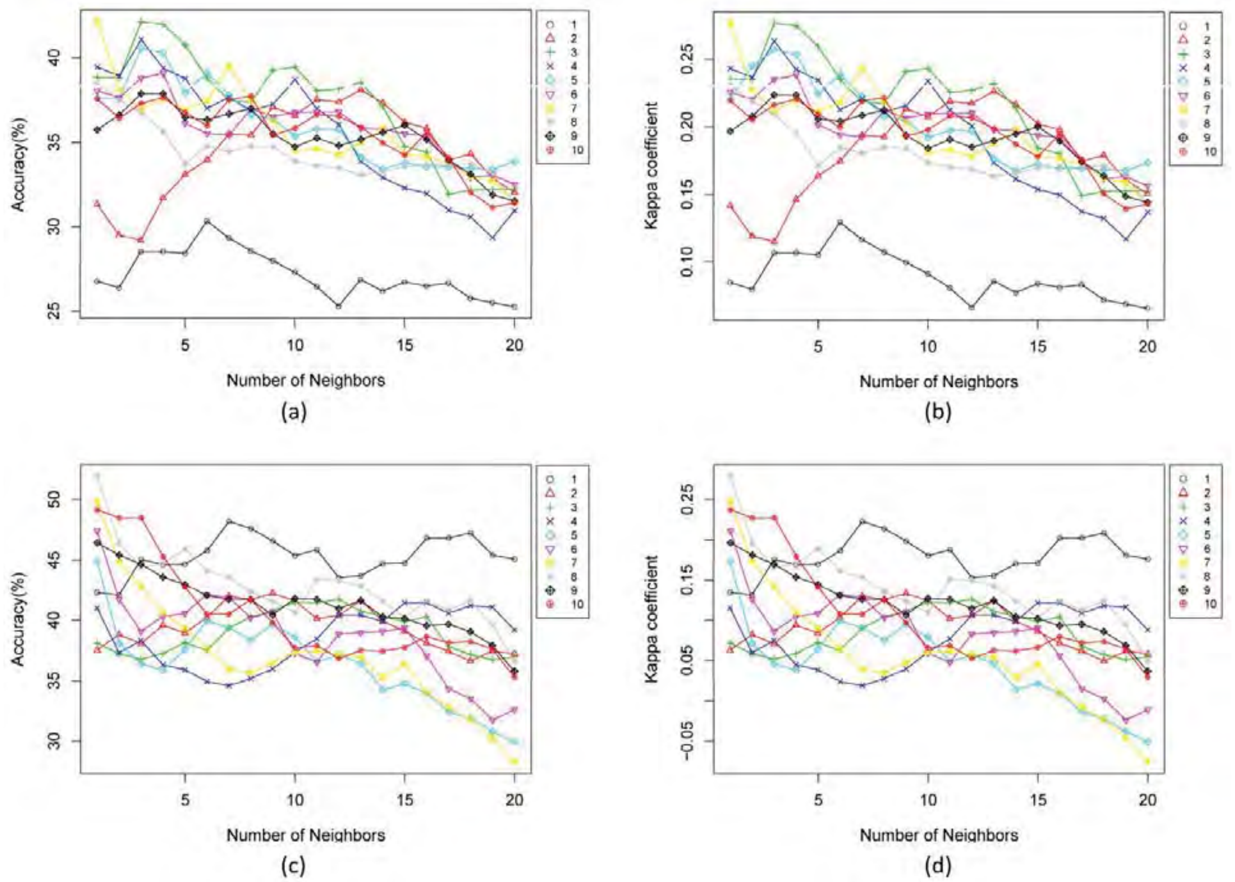


Figure 13. Average accuracy and kappa coefficient. (a), (b) grading for both *"Cancer"* and *"Benign"* samples. (c), (d) grading for *"Cancer"* samples. Each line depicts the accuracy and kappa coefficient values of the corresponding number of features.

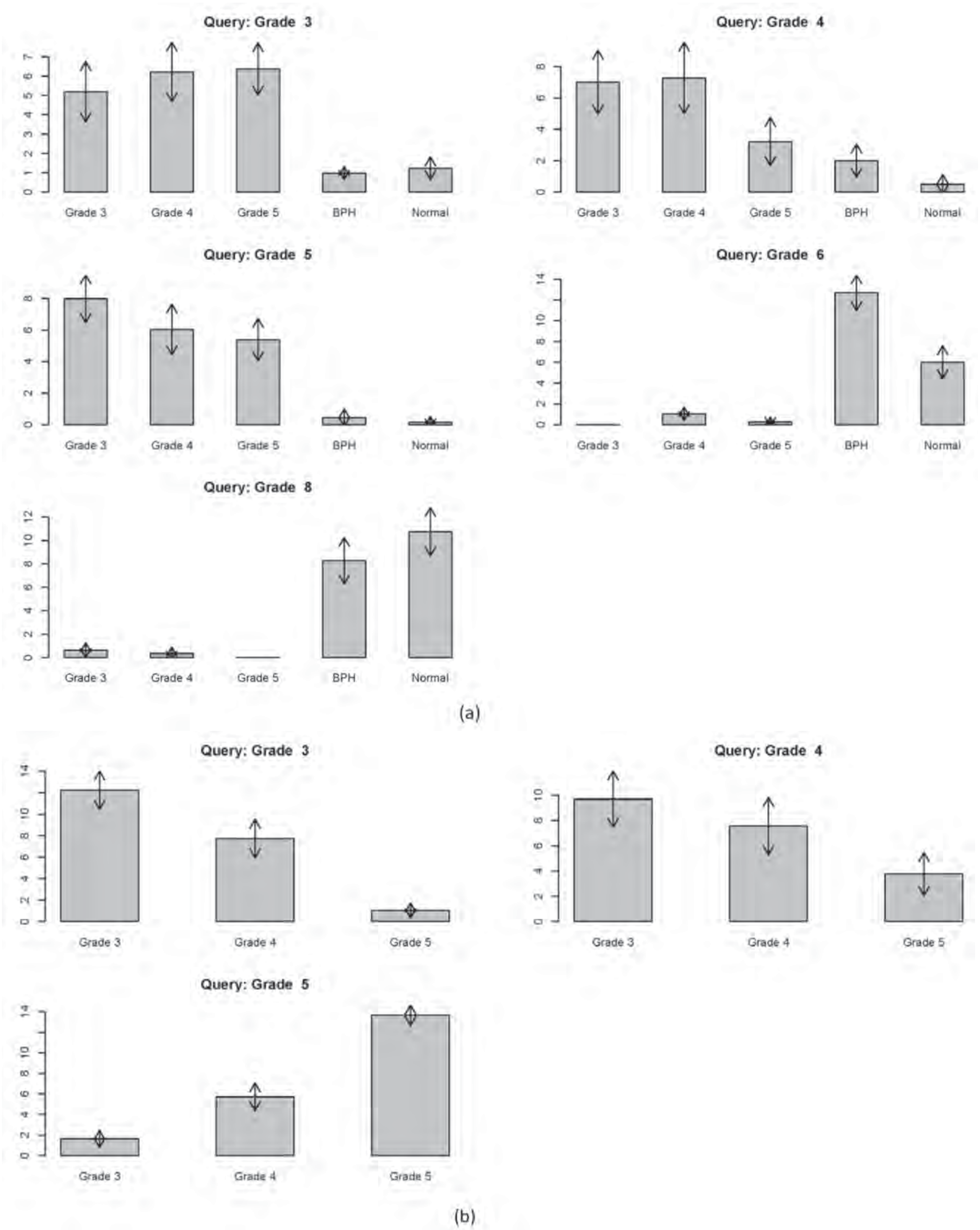


Figure 14. Distribution of the grade of the retrieved samples. (a) grading for both *"Cancer"* and *"Benign"* samples. (b) grading for *"Cancer"* samples. For the samples in each grade,

the grade of retrieved samples are counted and the average number of samples are shown. The arrows denote ±1 standard deviation of the number of samples.

Task: Develop protocols and validate Gleason grading of tumor (months 18-27)

The task above provides details of the development and LOO validation. More rigorous validations are needed but the preliminary results shows here have been used to validate the grading correspondence and the protocols we have developed, as noted above.

It is important to place the magnitude of our advance in context. Several research efforts have been made to develop automated systems for the grading of prostate tissues. The majority of systems have been used texture and/or morphological features to characterize and classify tissue samples into correct classes. However, the information which pathologists will obtain by using such methods may be limited since these only provide the predicted grade in general. The prediction also relies on the training data. Most importantly, these prior efforts always sought to match a sample completely to provide a diagnosis, rather than provide matching candidates. Further, the role of other modalities in the process was not clear. Here, we may also use IR chemical imaging data in matching. Our premise is that tissue samples which have the same grade and similar characteristics and patterns with the sample of interest will afford more information to pathologists and hence, the system enables a matching to a database rather than seeking to provide an unequivocal diagnosis.

Future outlook enabled by this progress:
The matching system would be implemented first for a clinical trial and then, would be ready for commercial translation. While a true clinical trial is the next step, some further development of the actual methods may be expected. We have built the method into existing software as a user-friendly software.

Task 3. Develop mathematical framework to correlate spectral, spatial and clinical parameters with cancer progression

a. Identify and validate spectral metrics and develop spatial metrics indicative of tumor grade (Months 27-30)
b. Develop prediction algorithm for predicting outcome (Months 30-36)

Activities:
We have imaged 460 patients with full outcome data and identified several metrics that are indicative of tumor grade (please see task 2 as well). A mathematical framework for correlate the spectral, spatial and clinical parameters with cancer progression has been built using logistic regression. The prediction algorithm is available for use and will be validated. The task for this project was to develop the algorithms and, hence, the task is complete.

Task 3. Develop mathematical framework to correlate spectral, spatial and clinical parameters with cancer progression

Goal: The goal of this task was to evaluate tissue with a view to predict outcome. The emphasis especially was on spectral features (metrics) that could be used.

a. Identify and validate spectral metrics and develop spatial metrics indicative of tumor grade (Months 27-30)
b. Develop prediction algorithm for predicting outcome (Months 30-36)

Activities:

We obtained a set of 460 samples in which patients were matched for age, PSA, grade and stage. Prostate cancer within half of the patients recurred within 5 years while the others were recurrence-free at 10 years. Samples from the entire data set were imaged. We then developed spectral metrics to characterize the samples. The relatively simple methods developed previously were not able to segment tissue into lethal (recur within 5 years) from indolent disease (does not recur in 5 years). Hence, we examined two avenues. In the first, we sought to examine if sources of variance were clouding the segmentation ability of the methods. Knowledge of the various sources of variance and their effects would help determine the mathematical model to be used. In the second, we sought to examine a major issue in the development of biomarkers – namely the difference in data sets and determine if the differences between data sets could be repaired by computational methods. We describe the first study first and the second one later.

Analysis of Variance in IR imaging data from prostate tissue

Most biomedical samples, including cells and tissue samples encountered in the prostate, are chemically complex and simple molecular compositions cannot be obtained. Hence, the analysis of complex tissues often relies on treating the IR spectrum as a signature of the identity and physiologic state. Many studies seek to find the spectral differences between given classes of samples from a statistical, rather than purely biochemical, perspective. These classes of samples may be different grades of disease or benign tissue, for example. Finding an analytical technique that can distinguish between disease states is of tremendous technological and medical importance as it can potentially aid clinicians and help prevent errors. IR imaging can potentially provide a solution by correlating spectral or spatial features with disease states. When suitable correlations between spectral differences and classes are found, a protocol may be constructed that allows for detection of these disease states in a practical application. Though conceptually straightforward, this approach is exceptionally challenging not only because of the subtle differences between various components and disease states in tissue but also because of the variation in IR spectra that may arise due to other factors and obscure differences between disease states. This variation in spectral differences overwhelming differences due to disease states is likely a primary cause for the failure of many analytical methods in providing robust protocols. Finally, the sample population under consideration may be of limited size, raising statistical issues in analyses and inferences. The analyses could be biased as the given samples may not be representative of the entire population. The latter two considerations can be addressed by careful study design and subsequent analysis. The question of analytic variability remains to be resolved and is a topic of much interest in infrared spectroscopy and other analytical technologies

Analytic variability can arise from (a) noise in signal measurement, (b) from differences within the tissue that leads to differences both within a given sample and between samples from the same patient, (c) differences between patients due to biologic diversity, (d) differences due to sample handling in different clinical settings or research groups and (e) due to causes not falling into any of the above categories. The variation may also be understood to be biological, technical or residual. Biological variation arises from different biological characteristics of samples such as patients, tissues, cells, subcellular components, etc. It is natural and expected variation, and often of interest in an experiment. Technical variation is attributable to both sample preparation and FT-IR imaging techniques. Potential sources of technical variation include tissue acquisition, fixation, and sectioning, placement of tissue section on the slide and post-preparation handling. The very process of data acquisition also introduces variation, such as measurement noise. Minimizing technical variation ensures data of high quality. Residual variation refers to the unexplained variation in the experiment; for example, environmental conditions – room temperature and humidity – that may not be part of the sample or acquisition characteristics. Although thoroughly identified, these potential sources of variation may never be complete. Accordingly, residual variation will be present and, on occasion, can have a substantial impact on the analyses. In such a case, we may either re-identify potential sources of variation or re-design the experiment.

Understanding the relative importance of each of these factors and explaining the variance observed in large scale tissue studies is critical for developing any real-world application. While an understanding of the contributions of variance by various sources can result in improved protocol designs, the lack of such understanding brings into question the performance of any developed protocol. Hence, in this manuscript, we develop a framework to understand analytic variability and its sources in infrared spectroscopic imaging of tissue. This understanding may be extended to other analytical techniques and imaging modalities, in general, and may be used to improve the practice of IR spectroscopic imaging for biomedical analysis in particular. The first challenge to understanding variability is to obtain a data set of sufficient diversity and size. Tissue microarrays (TMAs), to this end, are an excellent tool and have been used previously in a number of studies. TMAs consist of many samples of tissue arranged in a grid pattern. Multiple samples are usually included from the same person, a population of different people and, often, from different clinical settings is includes. Multiple TMAs may further be employed to increase sample set diversity and size. The effect of the various sources of variation can be analyzed by applying analysis of variance (ANOVA) model to the acquired data set. ANOVA is a popular statistical model for partitioning the total variance of the measured quantity in an experiment into various identifiable factors (or sources of variation), and has been applied for analyzing several spectroscopic imaging data: chemical compounds, collagen types, skin lesions, and plant species. However, to our knowledge, ANOVA has not been applied to spectroscopic imaging of tissue. Here, we present appropriate ANOVA models for different experimental designs of IR imaging data from TMAs, evaluate the statistical significance of the sources of variance, estimate variance contributions of the identified sources, and quantify the relative contributions of the sources to the total variation in the data. Finally, after examining the effect of the sources of variance, we also find the most discriminative spectral features and address the aspects of FT-IR imaging and TMA techniques that can be improved for better diagnostic protocols in prostate cancer.

Four experimental TMAs, containing prostate tissue samples, were obtained from different sources (Tissue microarray research program at the National Institutes of Health and Clinomics Inc.). The four TMAs contain respectively (i) 86 samples from 16 patients, (ii) 123 samples from 40 patients, (iii) 121 samples from 80 patients, and (iv) 240 samples from 180 patients. FT-IR-TMAs were taken at a spatial pixel size of 6.25 μm and a spectral resolution of 4 cm^{-1}. The spectral profile of a pixel spans a spectral range of 4,000–720 cm^{-1}. FT-IR data is converted into

93-dimensional data where each dimension corresponds to a spectral feature, which can be peak ratios, peak areas or peak centers of gravity. We note that the unit of observation in a spectral analysis is a pixel, but, in designing TMAs, the unit of interest is a tissue sample (called "core") or a patient. The number of pixels, especially of a single histological type such as epithelium, often varies substantially across cores and resulting data imbalance may greatly affect the results of the analysis. Therefore, we do not employ the entire collection of pixels (or cores) in TMAs, but address the issue of data imbalance by taking sub-samples of cores and sub-samples of pixels within each core in an attempt to balance the data for each group. The pixels corresponding to histologic classes were provided by either an automated histologic recognition method or a pathologic review.

Between-histologic class ANOVA model. In a typical TMA setting, many cores are placed in an array, one or more cores are obtained from a patient, and cores are often composed of multiple histologic classes such as epithelium, stroma, muscle, blood, and nerves, i.e., patients nested in an array, cores nested in a patient, and histologic classes nested in a core. Accordingly, variability in FT-IR-TMAs data is also distributed in a hierarchical fashion. Identifying five potential sources of variation (array, patient, core, histologic class and residual error), we present the following ANOVA model ("*between-histologic class model*") for this TMA design,

$$y_{ijklw} = \mu + \alpha_i + \beta_{j(i)} + \gamma_{k(j(i))} + \delta_l + \alpha\delta_{il} + \beta\delta_{j(i)l} + \gamma\delta_{k(j(i))l} + \omega_{ijklw} + \varepsilon_{ijklw} \qquad (1)$$

where *y* represent IR absorption of a pixel ($w = 1,...,n_\rho$) in a spectral feature of interest, *μ* is the overall mean, and *α*, *β*, *γ*, and *δ* denote array ($i = 1,...,n_\alpha$), patient ($j = 1,...,n_\beta$), core ($k = 1,...,n_\gamma$), and histologic class ($l = 1,...,n_\delta$) effect, respectively. *αδ*, *βδ*, and *γδ* are called interaction effects whereas *α*, *β*, *γ*, and *δ* are designated as main effects. ω_{ijklw} and ε_{ijklw} represent measurement error and residual error effects, respectively. On the contrary to the hierarchical structure of array, patient, and core effects, histologic class effect is crossed with each of array, patient, and core effects. Hence, this design is called a partly nested ANOVA. Since both fixed (histologic class) and random (array, patient, and core) factors present, it is also called a mixed effects ANOVA model (see Supporting Information for details).

The effect of the factors and their true variances can be estimated by computing ANOVA table and applying expected mean squared method which equates the observed and expected mean squares[18] (see Supporting Information for details). The total variance for (1) model can be written as $\sigma^2_{total} = \sigma^2_\alpha + \sigma^2_{\beta(\alpha)} + \sigma^2_{\gamma(\beta(\alpha))} + \sigma^2_\delta + \sigma^2_{\alpha\delta} + \sigma^2_{\beta\delta} + \sigma^2_{\gamma(\beta(\alpha))\delta} + \sigma^2_\omega + \sigma^2_\varepsilon$ where σ^2_α, $\sigma^2_{\beta(\alpha)}$, $\sigma^2_{\gamma(\beta(\alpha))}$, and σ^2_δ respectively indicate variance components of array, patient, core, and histologic class effects, and σ^2_ω and σ^2_ε are variance components of measurement error and residual error, respectively. σ^2_α, $\sigma^2_{\beta(\alpha)}$, $\sigma^2_{\gamma(\beta(\alpha))}$, and σ^2_δ can be attributable to biological variation as well as technical variation. These are due to biological variation because samples possess different biological characteristics. There is also technical variation in that variation can arise from any step in TMA preparation. σ^2_ω belongs to technical variation and is separately estimated on the assumption that it follows an independent and identically distributed Gaussian distribution over the entire spectral regions. We first compute the noise variance over the non-absorbing IR spectral regions (1900-2100cm^{-1}) and estimate measurement error for each spectral feature σ^2_ε is complex and reflects the combined effects of biological variation (pixel-to-pixel variation), technical variation (processing error), and other unexplained experimental variations. Hence, thorough inspection of residual error may be necessary for a precise and incisive analysis.

TMAs are often obtained from different sources, and the effect of the factors could differ significantly across TMAs. In order to further examine the differences, we estimate the variance components for each TMA by restricting the (1) ANOVA model to a single TMA. That is, we fitted IR data of each TMA to the following model,

$$y_{jklw} = \mu + \beta_j + \delta_l + \gamma\delta_{jl} + \gamma\delta_{k(j)l} + \omega_{jklw} + \varepsilon_{jklw}. \qquad (2)$$

Similarly, the total variance is $\sigma^2_{total} = \sigma^2_\delta + \sigma^2_\beta + \sigma^2_{\gamma(\beta)} + \sigma^2_{\beta\delta} + \sigma^2_{\gamma(\beta)\delta} + \sigma^2_\omega + \sigma^2_\varepsilon$. This model is also a partly nested ANOVA.

Between-array ANOVA model. Different histologic classes possess dissimilar chemical properties and cellular functions. It may introduce substantial variation to FT-IR-TMAs data. Eliminating histologic class factor and other related factors from the (1) model, we further examine the effect of histologic class on the data. The ANOVA model ("*between-array model*") can be expressed as

$$y_{ijkw} = \mu + \alpha_i + \beta_{j(i)} + \gamma_{k(j(i))} + \omega_{ijkw} + \varepsilon_{ijkw}. \qquad (3)$$

Since all factors are random, it is a nested random effects model. The total variance of data can be stated as $\sigma^2_{total} = \sigma^2_\alpha + \sigma^2_{\beta(\alpha)} + \sigma^2_{\gamma(\beta(\alpha))} + \sigma^2_\omega + \sigma^2_\varepsilon$. Since heterogeneous histologic classes are merged into a core, biological variation may increase in the model.

Between-subcellular component ANOVA model. The histologic classes are composed of a number of subcellular components (membrane, cytoplasm, nucleus, cytoskeleton, etc.). By further separating a histologic class into subcellular components, we can examine the effect of subcellular components for the histologic classes. The model is identical to the above (1) ANOVA model. The only difference is that the histologic class effect is replaced with subcellular component effect. Here, we restrict the model ("*between-subcellular component model*") to a single array as follows:

$$y_{kmw} = \mu + \gamma_k + \varphi_m + \gamma\varphi_{km} + \omega_{kmw} + \varepsilon_{kmw} \qquad (4)$$

where φ represents sub-cellular component ($m = 1, ..., n_m$) effect, which is fixed. This is a two factor crossed and mixed effects model, and the total variance is expressed as $\sigma^2_{total} = \sigma^2_\gamma + \sigma^2_\varphi + \sigma^2_{\gamma\varphi} + \sigma^2_\omega + \sigma^2_\varepsilon$. As did in *between-array model*, eliminating subcellular component factor and interaction effect between subcellular component and core from the (4) model, the following ANOVA model ("*within-epithelium model*") is constructed,

$$y_{kw} = \mu + \gamma_k + \omega_{kw} + \varepsilon_{kw}. \qquad (5)$$

This is a random effects model, and the total variance is expressed as $\sigma^2_{total} = \sigma^2_\gamma + \sigma^2_\omega + \sigma^2_\varepsilon$.

Variance component analysis identifies discriminative features for histologic analysis. Three TMAs (i, ii, iii) were used in this experiment. From each TMA, 26 sample cores from 13 patients were selected, and 200 pixels were chosen from each histologic class in a core. The histologic segmentation was conducted by a Bayesian classifier[19], built on 18 spectral features and achieved >0.99 AUC on cell type classification. Although several histologic classes present, we only consider epithelium and stroma; data imbalance in other classes is severer. Using *between-histologic class model*, ANOVA table (Table 1) and the portions of total variance due to the associated factors (Fig. 1A) were computed. 21 out of 93 features were dominated by histologic class effect, and either array effect or residual error introduced the most variation into the data other than the 21 features. The high variability in histologic class effect indicates that epithelium and stroma greatly differ in their IR absorption values, i.e., dissimilar chemical properties. Thus, the 21 features are capable of histologic analysis, and for the purpose of histologic discrimination, these features could serve as good candidates. In fact, 5 of them were included in the Bayesian cell type classifier. This may be attributable to the difference between the datasets or redundancy in the features. The Bayesian classifier was optimized for the classification, but variance components only show the ability of a single spectral feature, not their combined effects. It is probable that due to the redundancy the rest of the 21 features were not selected by the Bayesian classifier. Moreover, patient factor has very little effect on the total variation of the data, indicating that inferences made or models built on the data would be applicable to the entire patient population without or with very few restrictions or complication. Although its contribution to the total variance is small, larger variance from core effect than from patient effect suggests that the selection of cores is more important than that of patients in constructing TMAs. Since the size of a core is relatively small compared to the entire tissue or organ, it is likely that some of the selected cores are not representative of the tissue or organ. We also note that the small number of core samples could affect the variance estimates. We also note that interaction effects, by and large, were negligible except the interaction between core effect and histologic class effect. Interestingly, there were 19 features which were dominated by array effect, and these may need further assessment and reveal array-specific characteristics. Furthermore, we assessed the statistical significance of histologic class effect by computing *F*-test statistics, which is the ratio of the mean square of histologic class effect and the means square of the interaction effect of histologic class and array (see Supporting Information for details). Computing *p*-values for histologic-class effect, the lower p-values, the larger portions of explained variance were observed in general (Fig. 7A); the rank correlation coefficient of -0.57 (*p*-value≈0.0) was obtained between the portions of explained variance and the *p*-values. Thus, both variance components and *F*-test confirm the discriminative ability of the features for histologic analysis of tissue samples.

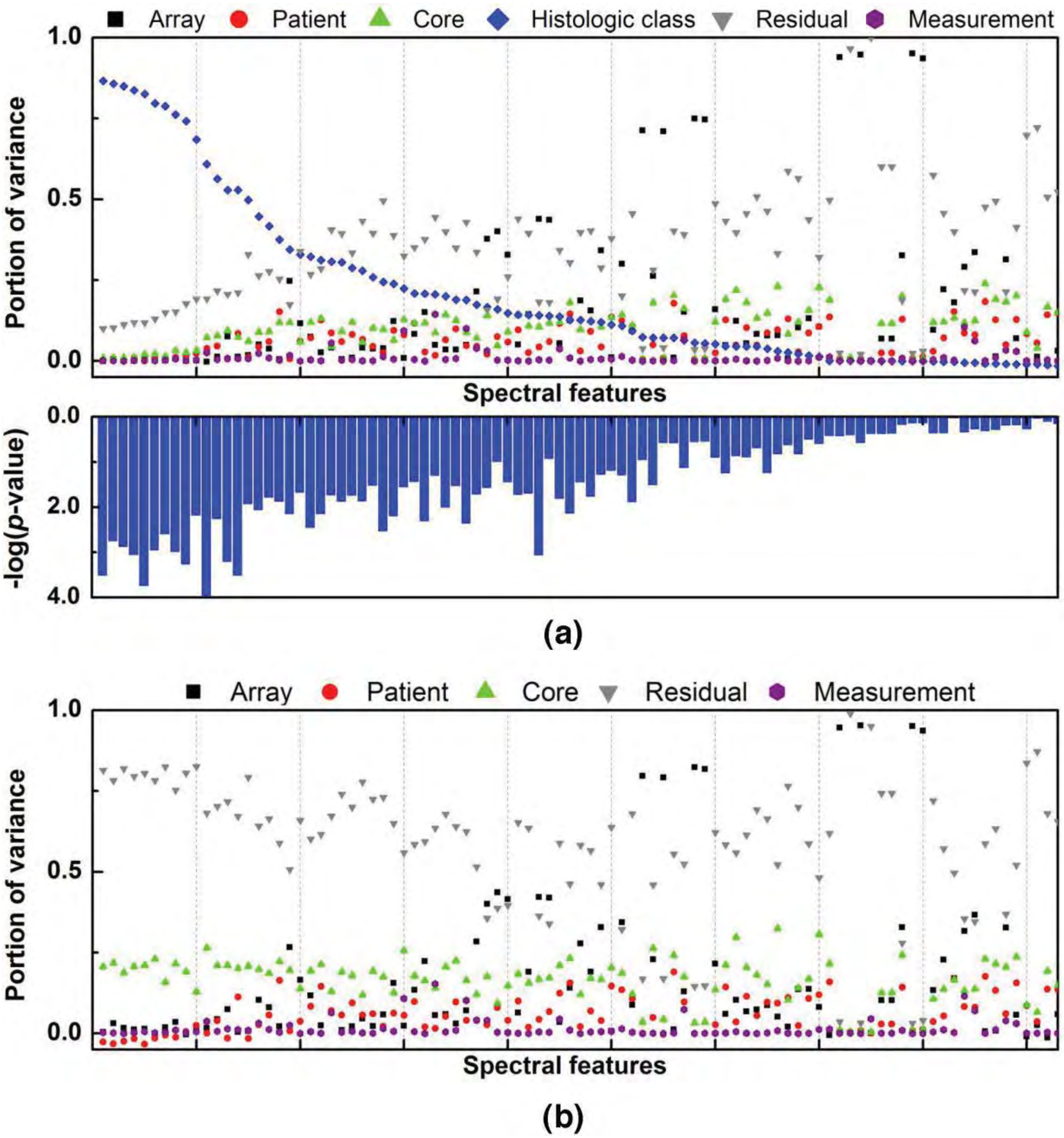


Figure 15. Portions of explained variance with and without histologic class factor. The portions of total variance explained by the associated factors are estimated for (a) *between-histologic class model* and (b) *between-array model* and plotted over 93 spectral features. (a) *p*-values of histologic class effect are shown at the bottom. (a)(b) The spectral features are ordered by the portion of total variance due to (a) histologic class effect. Interaction effects are not shown for (a).

Table 3. An example of ANOVA table for *between-histologic class model*.

Source	df	MS	EMS
Array	2	0.01678	$\sigma_{\omega}^{2}+\sigma_{\varepsilon}^{2}+400\sigma_{\gamma(\beta(\alpha))}^{2}+800\sigma_{\beta(\alpha)}^{2}+10400\sigma_{\alpha}^{2}$
Patient(Array)	36	0.02462	$\sigma_{\omega}^{2}+\sigma_{\varepsilon}^{2}+400\sigma_{\gamma(\beta(\alpha))}^{2}+800\sigma_{\beta(\alpha)}^{2}$
Core(Patient(Array))	39	0.01345	$\sigma_{\omega}^{2}+\sigma_{\varepsilon}^{2}+400\sigma_{\gamma(\beta(\alpha))}^{2}$
Histologic class	1	46.9907	$\sigma_{\omega}^{2}+\sigma_{\varepsilon}^{2}+200\sigma_{\gamma(\beta(\alpha))\delta}^{2}+400\sigma_{\beta\delta}^{2}+5200\sigma_{\alpha\delta}^{2}+15600\sigma_{\delta}^{2}$
Array*Histologic class	2	0.01180	$\sigma_{\omega}^{2}+\sigma_{\varepsilon}^{2}+200\sigma_{\gamma(\beta(\alpha))\delta}^{2}+400\sigma_{\beta\delta}^{2}+5200\sigma_{\alpha\delta}^{2}$
Patient(Array)*Histologic class	36	0.01604	$\sigma_{\omega}^{2}+\sigma_{\varepsilon}^{2}+200\sigma_{\gamma(\beta(\alpha))\delta}^{2}+400\sigma_{\beta\delta}^{2}$
Core(Patient(Array))*Histologic class	39	0.01356	$\sigma_{\omega}^{2}+\sigma_{\varepsilon}^{2}+200\sigma_{\gamma(\beta(\alpha))\delta}^{2}$
Residual error	31044	3.443e-4	σ_{ε}^{2}
Measurement error		4.601e-6	σ_{ω}^{2}

F-test statistics (Histologic class)	3980.5
p-value (Histologic class)	0.000251

df, MS and EMS denote degrees of freedom, mean squares, and expected mean squares, respectively. * indicates the interaction effect between factors.

Variance component analysis reveals weak discriminative feature for subcellular components analysis. To examine the effect of subcellular components, in the (iv) TMA, epithelial cells were further divided into two subcellular components: cytoplasm-rich and nucleus-rich. The cytoplasm-rich and nucleus-rich pixels were selected by a pathologic review. 40 cores from 40 patients were chosen, and 100 pixels for each of the two components were extracted to build *between-subcellular component model*. As shown in Fig. 2A, subcellular component effect is the dominant source of variation for only 9 features, and residual error is the most dominant factor for the rest of the features. Although subcellular component effect is the primary source of variation, the variance estimate of subcellular component effect does not overwhelm that of other effects as opposed to the huge differences between histologic effect and other effects in *between-histologic class model*. Thus, with the selected cytoplasm-rich and nucleus-rich pixels, we do not expect to observe a notable difference in IR spectra. This result is consistent with the previous work where ~0.72 AUC was obtained in classifying pixels into cytoplasm-rich and nucleus-rich pixels. Residual error is attributable to the similarity in the underlying chemical components, errors in selecting the cytoplasmic and nucleus pixels, and limitation in FT-IR imaging. That is, both biological and technical variations contribute to residual error. Biological variations could be reduced by minimizing cytoplasm and nucleus segmentation errors or obtaining higher-resolution FT-IR imaging. Performing repeated measurement of FT-IR imaging on the same TMAs could alleviate the contribution of technical variations. In addition, we computed *F*-test statistic for subcellular component effect as the ratio of the mean square of subcellular component effect and the means square of the interaction effect of subcellular component and core. Analogous to the results from *between-histologic class model*, the larger explained variance due to subcellular component effect, the lower *p*-values we observed, and the rank correlation coefficient of -0.33 (*p*-value>0.1) was obtained between the portions of explained variance and the *p*-values; however, the *p*-values were often too small (~0), misleading about the significance of subcellular component effect. Accordingly, *F*-test could not effectively provide discriminative features whereas variance components suggest weakly discriminating features. We also note that the features, owning high variation from histologic class effect in between-histologic class model, were not dominated by subcellular component effect in general. This indicates that the chemical properties to distinguish histologic classes differ from the properties to differentiate subcellular components.

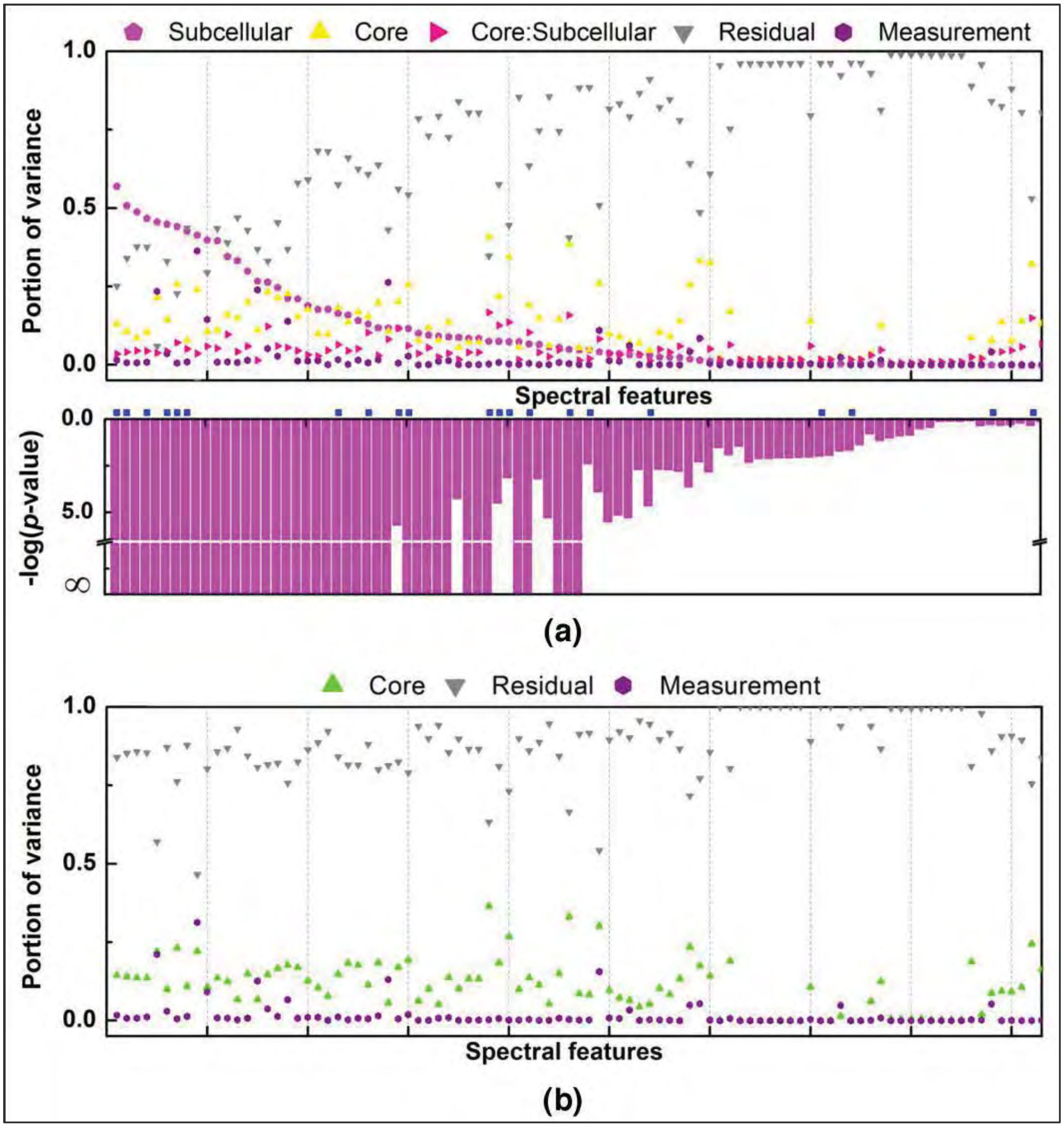


Figure 16. Portions of explained variance with and without subcellular component factor. The portions of total variance explained by the associated factors are estimated for (a) *between-subcellular component model* (b) *within-epithelium model* and plotted over 93 spectral features. (a) *p*-values of subcellular component effect are shown at the bottom, and the blue boxes indicate that the corresponding features are histologic class-dominant. (a)(b) The spectral features are ordered by the portion of total variance due to (a) subcellular component effect.

Biological variation is the main source of variation in residual error within a core and epithelium. Making no differentiation between histologic classes in a core, we fitted the FT-IR-TMAs data, used to build *between-histologic class model*, into *between-array model*, and the portions of total variance explained by each factor were computed. As shown in Fig. 7B, residual error is, in general, the dominant source of variation over the 93 features. The effects of four other factors (array, patient, core, and measurement error) were relatively small, and either array effect or core effect was mostly the second dominant source of variation. 16 features were dominated by array effect, of which 11 features were also array effect-dominant features in *between-histologic class model*. In comparison with *between-histologic class model*, combining histologic classes, we observed that residual error substantially increased in many features, especially for the 21 histologic class-dominant features. Similarly, we constructed the *within-epithelium model* using the data fitted into *between-subcellular component model*. Estimating the portions of variance due to core, measurement error, and residual error effects, residual error dominated over the other two effects in the entire 93 features (Fig. 8B). Compared to the variance components from *between-subcellular component model*, we again observed significant increase in residual error for numerous features including the 9 subcellular component-dominant features. Both histologic class and subcellular component factors group similar pixels in chemistry into the same group, as a result, decreasing biological variation. This leads us to conclude that biological variation is the main source of variation in residual error, especially for those 21 histologic class-dominant and 9 subcellular component-dominant features in the data. However, note that the interpretation of histologic class effect and subcellular component effect should be limited to the population under the experiment since both effects are fixed.

Differences in the effect of the associated factors are observed across TMAs. In order to investigate the differences in variance estimates across TMAs, each FT-IR-TMA data is fitted to the (2) ANOVA model. The proportions of variance estimates were, in general, very similar across TMAs, and, comparing to *between-histologic class model*, the similar trends were observed for the main effects; 16 out of the 21 histologic class-dominant features (*between-histologic class model*) showed high variability due to histologic class effect across all three TMAs; the rest of the features were mostly dominated by residual error across TMAs. Examining the 19 array-dominant features from *between-histologic class model*, we observed the differences in the variance components of not only histologic class effect but also other main and interaction effects across TMAs. In Fig. 9, for the first four features, although residual error was the most dominant source of variation, the relative orders of other factors varied greatly across TMAs, for example, histologic class effect and patient effect in the (i) TMA differ from the other 2 TMAs; the next four features showed unusually high variability in the (i) TMA and moderate dominance in the (iii) TMA from histologic class effect, but, in the (ii) TMA, the effect was not dominant or its contribution is close to residual error; examining the last 11 features, the differences in the portions of variance due to both main and interaction effects were also observed. For histologic analysis, these 19 array-dominant features may be avoided. The four features, in particular, introducing high variation from histologic class effect in the (i) TMA could be specific to the population represented by the (i) TMA, and thus may distract the histologic analysis and its translation into clinical practice. Computing *p*-values of histologic class effect, as observed in *between-histologic class model*, features with higher variance components possess lower p-values, but weaker correlations between them (rank correlation coefficients of -0.36 ~ -0.43) were observed. We note that the computation of *F*-test statistic is not identical to *between-histologic class model*. Here, the denominator is the mean square of the interaction effect between histologic class and patient.

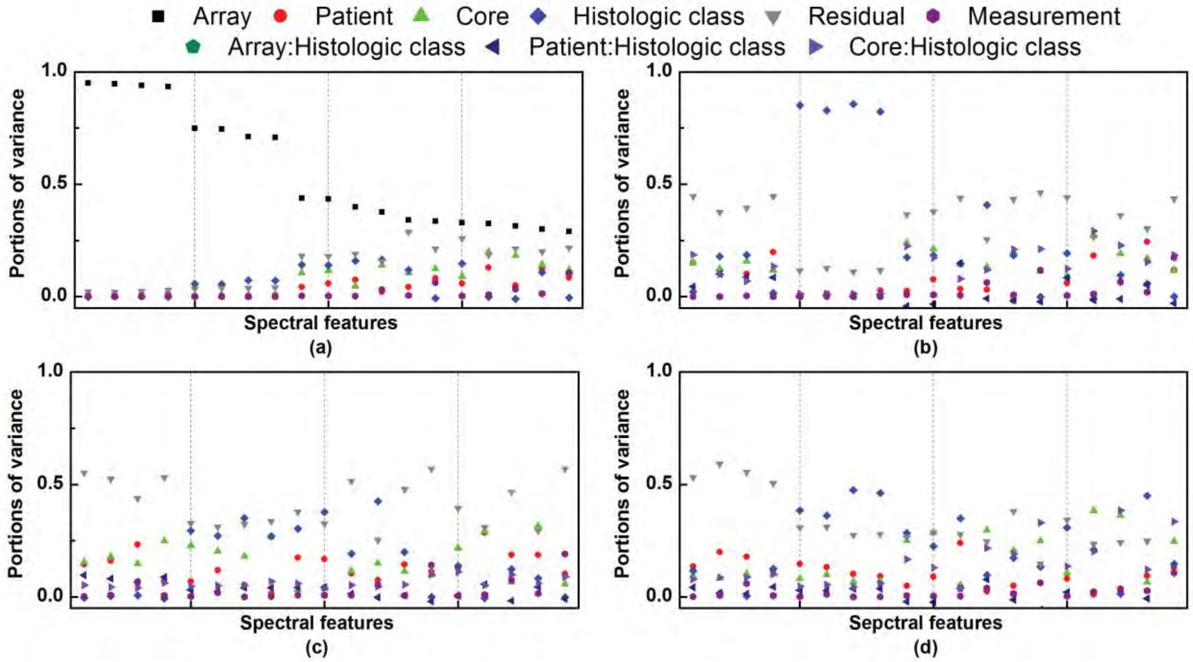


Figure 17. Portions of explained variance for array-dominant features across TMAs. Portions of variance are shown for (a) *between-array model* and (b,c,d) *between-histologic class model* restricted to each of (i,ii,iii) TMAs. Spectral features are ordered by the portion of total variance due to (a) array effect.

Correlating Changes Between two Data Sets

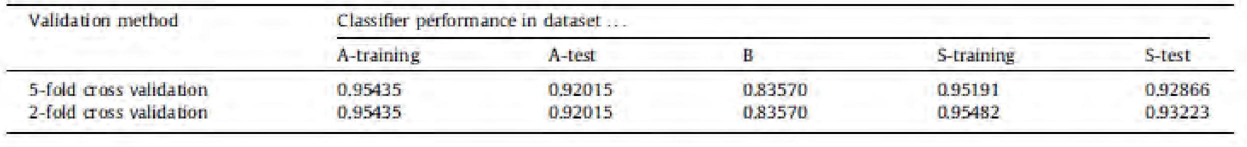
There is an underlying assumption on most model building processes: given a learned classifier, it should be usable to explain unseen data from the same given problem. Despite this seemingly reasonable assumption, when dealing with biological data it tends to fail; where classifiers built out of data generated using the same protocols in two different laboratories can lead to two different, non-interchangeable, classifiers. There are usually too many uncontrollable variables in the process of generating data in the lab and biological variations, and small differences can lead to very different data distributions, with a fracture between data. This paper presents a genetics-based machine learning approach that performs feature extraction on data from a lab to help increase the classification performance of an existing classifier that was built using the data from a different laboratory which uses the same protocols, while learning about the shape of the fractures between data that motivated the bad behavior. This is a critical step in understanding differences between our different prostate cancer data sets here.

The specific problem this study attempts to solve is the following: we have data from one laboratory (dataset A), and derive a classifier from it that can predict its category accurately. We are then presented with data from a second laboratory (dataset B). This second dataset is not accurately predicted by the classifier we had previously built due to a fracture between the data of both laboratories. We intend to find a transformation of dataset B (dataset S) where the classifier works. Evolutionary computing, as introduced by Holland, is based on the idea of the survival of the fittest, evoked by the natural evolutionary process. In genetic algorithms (GAs), solutions (genes) are more likely to reproduce the fitter they are, and random sporadic mutations help maintain population diversity. Genetic Programming (GP) is a development of those techniques, and follows a similar pattern to evolve tree-shaped solutions using variable-length chromosomes. Feature extraction ‘consists of the extraction a set of new features from the original features through some functional mapping’. Our approach to the problem can be seen as feature extraction, since we build a new set of features which are functions of the old ones. However, we have a different goal than that of classical feature extraction, since our intention is to fit a dataset to an already existing classifier, not to improve the performance of a future one. In this work, hence, we intend to demonstrate the use of GP-based feature extraction to unveil transformations in order to improve the accuracy of a previously built classifier, by performing feature extraction on a dataset where said classifier should, in principle, work; but where it does not perform accurately enough. We tested our algorithm first on artificially-built problems (where we apply ad hoc transformations to datasets from which a classifier has been built, and use the dataset resulting from those transformations as our problem dataset); and then on a real-world application where TMA data from two different medical laboratories regarding prostate cancer diagnosis are used as datasets A and B. Even though the method proposed does not attempt to reduce the number of features or instances in the dataset, it can still be regarded as a form of data reduction because it unifies the data distributions of two datasets; which results in the capability of applying the same classifier to both of them, instead of needing two different classifiers, one for each dataset.

In our previous work, we successfully applied a genetics-based approach to the development of a classifier that obtained human-competitive results based on FTIR data. However, the classifier built from the data obtained from one laboratory proved remarkably inaccurate when applied to classify data from a different hospital. Since all the experimental procedure was identical; using the same machine, measuring and post-processing; and having the exact same lab protocols, both for tissue extraction and staining; there was no factor that could explain this discrepancy. While one track was to understand the sources of variance, here we examined whether we could bridge the differences using GAs. The experimental and mathematical details are presented in the attached manuscript *“Repairing fractures between data using genetic programming-based feature extraction: A case study in cancer diagnosis”.*

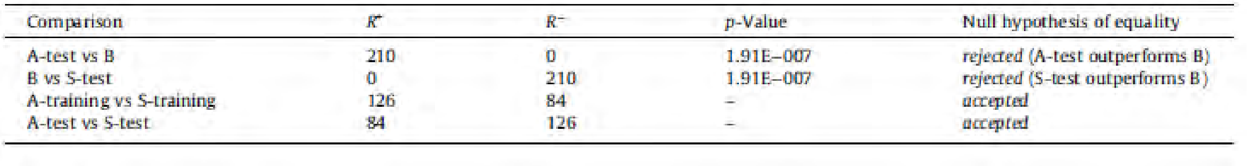
We summarize below the results for the prostate cancer problem in terms of classifier accuracy. The results obtained can be seen in Table 2. In that table, dataset A is the one from the first lab; which was used to build the classifier, dataset B is the one coming from the second lab, and dataset S is the result of the application of GP-RFD. To check whether the full dataset B was needed to evolve an effective transformation, we also tested using just half of it to train GP-RFD, and the other half to test (2-fold cross validation). These results are also included in Table 9. The performance results are excellent for a number of reasons. First and foremost, GP-RFD was able to find a transformation over the data from the second laboratory that made the classifier work just as well as it did on the data from the first lab, effectively finding the hidden perturbations that prevented the classifier from working accurately. The second positive conclusion to be obtained from the results is the generalization power of GP-RFD. As can be observed from the test results, GP-RFD does not ‘cheat’ by over-learning on the known data, and works well when transforming new, previously unseen, samples. Third, the results show GP-RFD was capable of obtaining excellent results using just half of the B dataset to train. This result highlights the power of the method to unveil the hidden transformation from a relatively low number of samples. We also performed a Wilcoxon signed-ranks test to evaluate the performance of GP-RFD over the case of study problem. In order to do it, we used the results from each partition in the 5-fold cross validation procedure. We ran the experiment four times, resulting in 4 5 = 20 performance samples to carry out the statistical test. As we did before, R+ corresponds to the first algorithm in the comparison winning, and R to the second one. Table 10 shows the results. The results on the case study problem are exactly the same as those achieved in the benchmark problems. We can then conclude GP-RFD was capable of repairing the existing fracture between the data from both laboratories. Again, this conclusion assumes class distribution did not change. It is a given in this case, since we know the class distribution to be equal in datasets A and B, but is an issue that has to be kept in mind when applying the method to other problems.

Table 4. Classifier performance results



Validation method	Classifier performance in dataset ...				
	A-training	A-test	B	S-training	S-test
5-fold cross validation	0.95435	0.92015	0.83570	0.95191	0.92866
2-fold cross validation	0.95435	0.92015	0.83570	0.95482	0.93223

Table 5. Statistical testing of the new protocol.



Comparison	R^+	R^-	*p*-Value	Null hypothesis of equality
A-test vs B	210	0	1.91E–007	*rejected* (A-test outperforms B)
B vs S-test	0	210	1.91E–007	*rejected* (S-test outperforms B)
A-training vs S-training	126	84	–	*accepted*
A-test vs S-test	84	126	–	*accepted*

We have presented GP-RFD, a new algorithm that approaches a common problem in real life for which not many solutions have been proposed in evolutionary computing. The problem in question is the repairing of fractures between data by adjusting the data itself, not the classifiers built from it. We have developed a solution to the problem by means of a GP-based algorithm that performs feature extraction on the problem dataset driven by the accuracy of the previously built classifier. We have tested GP-RFD on a set of artificial benchmark problems, where a problem dataset is fabricated by applying an ad-hoc disruption to an original dataset, and it has proved capable of solving all the transformations presented showing good performance both in train and, more importantly, test data. We have also being able to apply GP-RFD to the problem of prostate tissue classification, where data from two different laboratories regarding prostate cancer diagnosis was provided, and where the classifier learned from one did not perform well enough on the other. Our algorithm was capable of learning a transformation over the second dataset that made the classifier fit just as well as it did on the first one. The validation results with 5-fold cross validation also support the idea that the algorithm is obtaining good results; and has a strong generalization power. Lastly, we have applied a statistical analysis methodology that supports the claim that the classifier performance obtained on the solution dataset significantly outperforms the one obtained on the problem dataset. There is, however, one point where the proposed method has not been successful. The learned transformations have failed to provide any information about why the fracture appeared between the data from the two laboratories.

<u>Task 4. Write reports and finalize algorithms into software (Months 33-36)</u>

A number of reports (invention disclosure, conference etc.) have been written and manuscripts based on this work have been submitted and have been printed as detailed in the following sections.

In summary, the promised work has been accomplished to a reasonable degree and has opened up significant doors to future progress in prostate pathology as a research direction as well as for patients and clinicians.

Key Research Accomplishments

- A genetic algorithm based method to distinguish benign from malignant epithelium using infrared spectroscopic imaging data was shown to be effective. Large scale validation shows promising results and a manuscript is being written.
- We determined that one of the key factors in understanding our data was the spatial structure of the tissue, that closely affected the IR data. A series of simulations were conducted after developing a rigorous optical model to predict distortions. Results are reported in two manuscripts in *Anal. Chem.*
- A combination of IR and conventional pathology imaging has been developed and extensively validated.
- A method to correlate Gleason grades with measured data has been developed. Larger validation studies are needed.
- A number of patent applications and invention disclosures as well as peer-reviewed publications have resulted from these activities.

Reportable Outcomes……………………………………………………………

Manuscripts

Peer reviewed manuscripts published

1. M.J. Walsh, R.K. Reddy, **R. Bhargava** "Label-free Biomedical Imaging with Mid-Infrared Microspectroscopy" submitted (2011)
2. R.K. Reddy, M.J. Walsh, **R. Bhargava** "A framework for visualizing multidimensional relationships in large spectroscopic imaging data sets" *Appl. Spectrosc.* Submitted (2011)
3. M.J. Walsh, M.V. Schulmerich, **R. Bhargava** "Progress, critical challenges and a roadmap for translating infrared spectroscopic imaging for cancer histopathology" *Chem. Rev.*, To be submitted (2011 – Invited)
4. S. Holton, M.J. Walsh, A. Kajdacsy-Balla, **R. Bhargava** "Label-free characterization of cancer-activated fibroblasts in cell culture and clinical tissues using infrared spectroscopic imaging" *Biophys. J.* In press (2011)
5. S.E. Holton, M.J. Walsh, **R. Bhargava** "Localizing subcellular biochemical transformations in cancer-activated fibroblasts using high-resolution infrared spectroscopic imaging" *Analyst* **136**, 2953-2958 (2011) DOI: 10.1039/c1an15112f
6. R. Kong, **R. Bhargava** "Characterization of Porcine Skin as a Model for Human Transdermal Diffusion" *Analyst* **136**, 2359 - 2366 (2011) DOI: 10.1039/C1AN15111H
7. M.J. Nasse, M.J. Walsh, E.C. Mattson, R. Reininger, A. Kajdacsy-Balla,V. Macias, **R. Bhargava**,* C.J. Hirschmugl* "High-resolution Fourier transform infrared chemical imaging with multiple synchrotron beams" *Nat. Methods,* **8**, 413-416 (2011) DOI:10.1038/NMETH.1585 *[*Joint corresponding authors]*
   - News and Views: F.L. Martin "Shining a new light into molecular workings" *Nat. Methods,* **8**, 385–387 (2011)
   - Author profile by M. Baker *Nat. Methods,* **8**, 363 (2011)
8. J. G. Moreno-Torres, X. Llora, D.E. Goldberg, **R. Bhargava** "Repairing Fractures between Data using Genetic Programming-based Feature Extraction: A Case Study in Cancer Diagnosis" *Information Sciences*, In press (2011).
9. J.T. Kwak, S.M. Hewitt, S. Sinha, **R. Bhargava** "Multimodal microscopy for automated histologic analysis of prostate cancer" *BMC Cancer* **11**, 62- (2011) [Designated "Highly Accessed"]
10. B.J. Davis, P.S. Carney, **R. Bhargava** "Infrared Microspectroscopy of Intact Fibers" *Anal. Chem.* **83**, 525–532. (2011)

11. B. Kwon, C. Wang, K. Park, **R. Bhargava**, W.P. King "Thermomechanical Sensitivity of Microcantilevers in the Mid-infrared Spectral Region", *Nano Micro Thermophys Eng*, **15**, 16-28 (2011)
12. J. G. Moreno-Torres, X. Llora, D.E. Goldberg, **R. Bhargava** "On the homogenization of data from two laboratories using genetic programming" *Lec. Notes Comp. Sci.,* **6471/2010**, 185-197 (2010), DOI: 10.1007/978-3-642-17508-4_12.
13. A.K. Kodali, M.V. Schulmerich, R. Palekar, X. Llora, **R. Bhargava** "Optimized Nanospherical Layered Alternating Metal-dielectric Probes for Optical Sensing" *Opt. Exp.*, **18**, 23302-23313 (2010)
14. R.K. Reddy, **R. Bhargava** "Automated noise reduction for accurate classification of tissue from low signal-to-noise ratio imaging data" *Analyst*, **135**, 2818-2825 (2010) DOI: 10.1039/C0AN00350F
15. A.K. Kodali, X. Llora, **R. Bhargava** "Optimally tailored Raman spectroscopic probes for ultrasensitive and highly multiplexed assays" *Proc. Natl. Acad. Sci.*, **107**, 13620-13625 (2010) DOI: 10.1073/pnas.1003926107
16. A.K. Kodali, M.V. Schulmerich, J. Ip, G. Yen, B.T. Cunningham, **R. Bhargava** "Narrowband Mid-infrared reflectance filters using guided mode resonance" *Anal. Chem.*, **82**, 5697–5706 (2010) DOI: 10.1021/ac1007128
17. M.V. Schulmerich, A.K. Kodali, R.K. Reddy, L.J. Elgass, **R. Bhargava** "Dark field Raman Microscopy" *Anal. Chem.*, **82**, 6273–6280 (2010) DOI: 10.1021/ac1014194
18. R. Kong, R.K. Reddy, **R. Bhargava** "Characterization of Tumor Progression in Engineered Tissue using Infrared Spectroscopic Imaging" *Analyst* **135**, 1569-1578 (2010) DOI: 10.1039/c0an00112k
19. B.J. Davis, P.S. Carney, **R. Bhargava** "Theory of mid-infrared absorption microspectroscopy. II. Heterogeneous samples" *Anal. Chem.*, **82**, 3487–3499 (2010) DOI: 10.1021/ac902068e
20. B.J. Davis, P.S. Carney, **R. Bhargava** "Theory of mid-infrared absorption microspectroscopy. I. Homogeneous samples" *Anal. Chem.*, **82**, 3474–3486 (2010) DOI: 10.1021/ac902067p
21. X. Llora, A.Priya, **R. Bhargava** "Observer-Invariant Histopathology using Genetics-Based Machine Learning" *Nat. Computing*, **8**, 101-120 (2009)

Book Chapters

1. M.J. Walsh, **R. Bhargava** "Infrared spectroscopic imaging: an integrative approach to pathology" G. Popescu, ed. "Nanobiophotonics" McGraw-Hill (2010)
2. R.K. Reddy, **R. Bhargava** "Chemometric methods for biomedical Raman spectroscopy and imaging" M.D. Morris, P.Matousek, eds. "Emerging Raman Applications and Techniques in Biomedical and Pharmaceutical Fields", Springer-Verlag, Berlin Heidelberg (2010)
3. A.K. Kodali, **R. Bhargava** "Nanostructured Probes to Enhance Optical and Vibrational Spectroscopic Imaging for Biomedical Applications", Y.Y. Fu and A. Narlikar, eds. "The Oxford handbook of Nanoscience and Technology: Vol. III", Oxford University Press, Oxford, UK (2010)

Other manuscripts

1. M.J. Walsh, M.J. Nasse, F.N. Pounder, V. Macias, A. Kajdacsy-Balla, C. Hirschmugl, **R. Bhargava** "Synchrotron FTIR Imaging For The Identification Of Cell Types Within Human Tissues" AIP Conference Proc. Vol. 1214 WIRMS 2009 5th international workshop on infrared microscopy and spectroscopy with accelerator based sources pp. 105-107
2. F.N. Pounder, R.K. Reddy, M.J. Walsh, **R. Bhargava** "Validating the cancer diagnosis potential of mid-infrared spectroscopic imaging" Proc. SPIE 7186, art. no. 71860f
3. **R. Bhargava** , B.J. Davis, "Histologic models for optical tomography and spectroscopy of tissues" Proc. SPIE 7174, 71742H (2009), DOI:10.1117/12.810119

Presentations

Invited conference presentations

First author is the presenting author; First author is also the invited author unless indicated by *

1. **R. Bhargava**, M.J. Walsh, R.K. Reddy, J.T. Kwak, A. Balla "Chemical imaging for histopathology" New Frontiers and Grand Challenges in Laser-Based Biological Microscopy, Telluride, CO, August 2011
2. **R. Bhargava** "Infrared spectroscopic imaging for label-free biomedical informatics" OSA topical meeting on optical sensors, Toronto, June 2011
3. **R. Bhargava** "Infrared spectroscopic imaging: a new direction for an old chemical imaging technique" Central Regional Meeting of the ACS, Indiana, June 2011
4. **R. Bhargava** "Chemical imaging for histopathology", Pittcon 2011, Atlanta, March 2011
5. **R. Bhargava** "Systems pathology with infrared spectroscopic imaging" Pacifichem 2010, Honolulu, HI, December 2010
6. **R. Bhargava**, M.V. Schulmerich, A.K. Kodali, R.K. Reddy and M.J. Walsh "Chemical imaging for automated histopathology", Eastern Analytical Symposium, Somerset, November 2010
7. **R. Bhargava**, M.V. Schulmerich, A.K. Kodali, X. Llora, R.K. Reddy, M. Kole "Using computational modeling to improve Biomedical Raman microscopy", Eastern Analytical Symposium, Somerset, November 2010
8. **R. Bhargava** "Chemical imaging for molecular pathology" Cancer Colloquia VII, St. Andrews, Scotland, November 2010
9. **R. Bhargava**, R.K. Reddy, J. Ip, F.N. Pounder, M.V. Schulmerich, D. Mayerich, X. Llora, R. Kong, M.J. Walsh "Infrared Spectroscopic Imaging for Label-Free and Automated Histopathology" Frontiers in Optics, Rochester, October 2010
10. **R. Bhargava,** A. K. Kodali, M. V. Schulmerich, X. Llora and R. K. Reddy "Integrating physics with chemometrics for enhanced vibrational spectroscopic imaging", FACSS 10, Raleigh, October 2010 [Meggers award symposium]
11. **R. Bhargava,** M. V. Schulmerich and R. K. Reddy "Discrete frequency infrared spectroscopic imaging with a quantum cascade laser – rationale and potential", FACSS 10, Raleigh, October 2010
12. **R. Bhargava**, P.S. Carney, R.K. Reddy, A.K. Kodali "Modeling distortions in infrared spectroscopic imaging", FACSS 10, Rayleigh, October 2010
13. **R. Bhargava** "Enabling prostate pathology with infrared spectroscopic imaging – a roadmap for clinical translation", SPEC2010, Manchester, June 2010 (*Plenary opening lecture*)
14. **R. Bhargava** "Non-perturbing cancer diagnostics using infrared spectroscopic imaging", Pittcon 2010, Orlando, March 2010
15. **R. Bhargava** "Progress towards cancer pathology using infrared spectroscopic imaging" , Pittcon 2010, Orlando, March 2010
16. **R. Bhargava** "Enabling systems pathology by infrared spectroscopic imaging", Pittcon 2010, Orlando, March 2010
17. **R. Bhargava** "Pathology without pathologists?" Pathological Society of Great Britain and Ireland, London, January 2010
18. **Bhargava**, R.K. Reddy, M. Schulmerich, A.K. Kodali, F.N. Pounder B.J. Davis "Next-generation infrared imaging for biomedical spectroscopy", FACSS 09, Louisville, October 2009
19. **R. Bhargava**, J. Ip, A.K. Kodali, F.N. Pounder B.J. Davis "Ultrafast IR imaging for Biomedical applications", ICAVS-5, Melbourne, July 2009 (*Plenary Lecture*)
20. **R. Bhargava** "Imaging: Does it really offer more than 'just' pretty pictures", SAS 50 years symposium, Pittcon 09, Chicago, March 2009
21. **R. Bhargava**, R.K. Reddy "The critical role of controlled quality of spectral information and sampling on automated histologic recognition", Pittcon 09, Chicago, March 2009
22. **R. Bhargava,** F.N. Pounder, X. Llora and R.K. Reddy "Enhancing the tissue segmentation capability of fast infrared spectroscopic imaging via chemometric methods", FACSS08, Reno, September 2008
23. **R. Bhargava**, F.N. Keith, R.K. Reddy and A.K. Kodali "Practical infrared spectroscopic imaging instrumentation for translating laboratory results to clinical settings", FACSS08, Reno, September 2008

24. **R. Bhargava** "Spectroscopic Imaging for an Automated Approach to Histopathologic Recognition in Prostate Tissue" 82nd Annual North Central Section American Urological Association Meeting, Chicago, September 2008
25. **R. Bhargava**, R.K. Reddy, A.K. Kodali "Ultrafast mid-infrared spectroscopic imaging by combined computational and experimental optimizations" ISSSR 2008, Hoboken, June 2008
26. **R. Bhargava**, R.K. Reddy, R. Kong, G. Srinivasan "Engineering practical protocols for histopathology of human tissues and models using infrared spectroscopic imaging", Pittcon08, New Orleans, March 2008
27. **R. Bhargava**, R.K. Reddy, R. Kong, F. N. Keith, G. Srinivasan "Automated Cancer Histopathology by Practical Infrared Spectroscopic Imaging: Progress and Potential" The International Conference on Perspectives in Vibrational Spectroscopy (ICOPVS), Thiruvananthapuram, Kerala , India, February 2008 (*Plenary Lecture*)

Other invited presentations

28. National Institute for Standards and Technology (NIST), Gaithersburg, 2011
29. Center for nanoscale science and technology annual symposium, University of Illinois, Urbana, 2011
30. Young Breast Cancer Survivors coalition symposium, Urbana, 2010
31. Synchrotron Research Center, Madison, WI, 2010
32. iOptics Seminar, University of Illinois at Urbana-Champaign, Urbana, 2010
33. Bruker Optics users meeting, Boston, 2010
34. University of Illinois Cancer Center, UIC, Chicago, 2010
35. Department of Bioengineering, Ohio State University, 2009
36. Beckman Institute Director's Seminar Series, UIUC 2009
37. Department of Chemistry, University of Tennessee, Knoxville, 2009
38. BioInterest Group Seminar, Mechanical Science and Engineering, UIUC, 2008
39. Lester Wolfe Workshop, MIT, 2008
40. Translational Biomedical Research Seminar, Veterinary Medicine, UIUC, 2008
41. Vistakon, A Division of Johnson and Johnson, Jacksonville, 2008
42. Laser Science Center, Indian Institute of Technology, Kanpur, 2008

Contributed presentations

*First author is the presenting author, unless indicated by **

1. M.J. Walsh, D. Mayerich, E.L. Wiley, R. Emmadi, A. Kajdacsy-Balla, R. Bhargava R. "Mid-Infrared Spectroscopic Imaging for Breast Tissue Histopathology: Towards 'Stainless Staining" 1st Congress of the International Academy of Digital Pathology, Quebec, Canada, August 2011.
2. R.K. Reddy, B.J. Davis, P.S. Carney, **R. Bhargava** "Modeling Fourier transform infrared spectroscopic imaging of Prostate and breast cancer tissue specimens" IEEE International Symposium on Biomedical Imaging (ISBI), Chicago, March 2011
3. J.T. Kwak, S. Sinha, **R. Bhargava** "Histological segmentation for infrared spectroscopic imaging using frequent pattern mining" IEEE International Symposium on Biomedical Imaging (ISBI), Chicago, March 2011
4. M.J. Walsh, **R. Bhargava** "Towards Comprehensive Histopathological Analyses in Breast and Prostate Tissue Using Mid-IR Spectroscopic Imaging" FACSS 2010, Raleigh, October 2010
5. R.K. Reddy, B.J. Davis, **R. Bhargava** "Enhanced Models for Fourier Transform Infrared (FT-IR) Spectroscopic Imaging of Human Tissue Specimens" FACSS 2010, Raleigh, October 2010
6. M.J. Walsh, J. Ip, C. Cvetkovic, **R. Bhargava** "Mid-IR Imaging for Identification of Cells and Mucin Subtype in the Gastrointestinal Tract" FACSS 2010, Raleigh, October 2010

7. B. Kwon, M.V. Schulmerich, L. Elgass, R. Kong, S. Holton, **R. Bhargava**, W.P. King "Infrared Imaging Spectrometry using an Atomic Force Microscope" MRS Fall Meeting, Boston, November 2010
8. M.J. Walsh, **R. Bhargava** "Histopathological Analyses in Breast and Prostate Tissue Using Mid-IR Spectroscopic Imaging" SPEC 2010, Manchester, June 2010
9. R.K. Reddy, **R. Bhargava** "Modeling, Data Visualization and Histopathology using Fourier Transform Infrared (FT-IR) Spectroscopic Imaging of Human Tissue Specimens", BMES 2009 Pittsburgh, PA, October 2009
10. M. J. Walsh, M. J. Nasse, F. N. Pounder, V. Macias, A. Kajdacsy-Balla, C. Hirschmugl, **R. Bhargava** "Mid-infrared spectroscopic imaging of prostate tissue towards cancer diagnosis and prognosis", BMES 2009, Pittsburgh, October 2009
11. M. J. Walsh, M. J. Nasse, F. N. Pounder, V. Macias, A. Kajdacsy-Balla, C. Hirschmugl, **R. Bhargava** "Synchrotron FT-IR imaging for identification of cell types within human tissues", WIRMS 2009, Banff, Canada, September 2009
12. F.N. Pounder, **R. Bhargava** "Human-Competitive Histologic Follow-up to Breast Cancer Screening with Mid-IR Spectroscopic Imaging," Pittcon, Chicago, March 2009
13. R.K. Reddy, **R Bhargava** "Automated and fast histologic characterization in urology: progress towards an unmet clinical need", Urology: Diagnostics, Therapeutics, Robotics, Minimally Invasive, and Photodynamic Therapy, BiOS 2009, San Jose, CA
14. R.K. Reddy, F.N. Pounder, **R. Bhargava** "Validating the cancer diagnosis potential of mid-infrared spectroscopic imaging", SPIE Photonics West - BiOS 2009, San Jose, CA
15. J. Ip, **R. Bhargava** "Integrating instrumentation, computation and sampling for a high throughput approach to automated histology by mid-infrared microscopy", Advanced Biomedical and Clinical Diagnostic Systems VII, SPIE Photonics West - BiOS 2009, San Jose, CA
16. M.J. Walsh, F.N. Pounder, **R. Bhargava** "Spectral pathology in breast cancer using mid-infrared spectroscopic imaging", Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues VII, SPIE Photonics West - BiOS 2009, San Jose, CA
17. R. Bhargava, A.K. Kodali, F.N. Pounder, R.K. Reddy "High-speed Infrared Spectroscopic Imaging for Tissue Histopathology", EAS 2008, Somerset, November 2008

Funding received for based on work supported by this award

Project/Proposal Title:
Infrared microscopy for prostate pathology (Role: PI)
Source of Support: National Institutes of Health
Total Award Amount: $ 1 832, 819 Total Award Period Covered: 02/01/2010-12/31/2014
Location of Project: Urbana, IL
Person-Months Per Year Committed to the Cal: Acad: Sumr: 1.0

Funding applied for based on work supported by this award

None at present

Employment or research opportunities applied for and/or received based on experience/training supported by this award.

Dr. Brynmor Davis, a post-doctoral fellow working on this project obtained employment with Creare Inc., NH.

Dr. Gokulakrishnan Srinivasan, a post-doctoral fellow working on this project obtained employment with Bruker Optics.

Conclusion……………………………………………………………………………

The work accomplished demonstrates clear potential and protocols for classifying prostate tissue. If the protocols are validated in on-going larger studies and translated to the clinic, a new tool for prostate histopathology will be available for pathologists and benefits will be realized by patients.

So What Section

An automated method to assist prostate pathologists is available and can rapidly determine the presence of cancer in biopsies. An automated aid to grading is available to aid pathologists in making accurate decisions. Clinical translation of these discoveries can directly improve prostate healthcare, resulting in better treatment of individuals.

References…………………………………………………………………………….

[1] A Jemal, R Siegel, E Ward, T Murray, J Xu, C Smigal, MJ Thun Cancer statistics, 2006 *CA Cancer J Clin* **56**, 106-130 (2006).

[2] SM Gilbert, CB Cavallo, H Kahane, FC Lowe Evidence suggesting PSA cutpoint of 2.5 ng/mL for prompting prostate biopsy: Review of 36,316 biopsies. *Urology* **65**, 549-553 (2005).

[3] PF Pinsky, GL Andriole, BS Kramer, RB Hayes, PC Prorok, JK Gohagan, Prostate, Lung, Colorectal and Ovarian Project Team Prostate Biopsy Following a Positive Screen in the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial *J Urol* **173**, 746-750 (2005). discussion 750-751.

[4] PA Humphrey *Prostate Pathology* American Society of Clinical Pathology, Chicago (2003).

[5] EN Lewis, PJ Treado, RC Reeder, GM Story, AE Dowrey, C Marcott, IW Levin Fourier transform spectroscopic imaging using an infrared focal-plane array detector *Anal. Chem.* **67**, 3377-3384 (1995).

[6] R Bhargava, SQ Wang, JL Koenig Processing FTIR Imaging Data for Morphology Visualization *Appl Spectrosc* **54**, 1690-1706 (2000).

[7] DC Fernandez, R Bhargava, SM Hewitt, IW Levin Infrared spectroscopic imaging for histopathologic recognition *Nat. Biotechnol.* **23**, 469-474 (2005).

[8] Snively C M, Koenig J L. Characterizing the performance of a fast FT-IR imaging spectrometer. *Appl. Spectrosc.* 1999; **53** :170-177.

[9] Bhargava R, Levin, I W. Fourier transform infrared imaging: Theory and practice. *Anal. Chem.* 2001; **73** :5157 -5167.

[10] Bhargava R, Rebar T, Koenig J. Towards faster FT-IR imaging by reducing noise. *Appl. Spectrosc.* 1999; **53** :1313–1322.

[11] Bhargava R, Wang S, Koenig J. Route to higher fidelity FT-IR imaging. *Appl. Spectrosc.* 2000; **54** :486–495.

[12] Snively C, Katzenberger S, Oskarsdottir G, Lauterbach. Fourier-transform infrared imaging using a rapid-scan spectrometer. *Opt. Lett.* 1999; **24** :1841–1843.

[13] Green A, Berman M, Switzer P, Craig, M. A Transformation For Ordering Multispectral Data In Terms Of Image Quality With Implications For Noise Removal. *IEEE T. Geosci. Remote* 1988; **26** :65–74.

[14] Boardman J, Kruse F. Automated spectral analysis: a geological example using AVIRIS data, north Grapevine Mountains, Nevada. *Proceedings, ERIM Tenth Thematic Conference on Geologic Remote Sensing* 1994; 407–418.

[15] Wentzell P, Andrews D, Hamilton D, Faber K, Kowalski B. Maximum likelihood principal component analysis. *J. Chemometr.* 1997; **11** :339–366.

[16] Qin, S, Dunia, R. Determining the number of principal components for best reconstruction. *J. Process Contr.* 2000; **10** :245–250.

[17] Cattell R B. The Screen Test For The Number Of Factors 1. *Multivar. Behav. Res.* 1966; **1** :245–276.

[18] Wold S. Cross-Validatory Estimation Of Number Of Components In Factor And Principal Components Models. *Technometrics* 1978; **20** :397–405.

[19] Valle, S, Li,W, Qin, S. Selection of the number of principal components: The variance of the reconstruction error criterion with a comparison to other methods. *Ind. Eng. Chem. Res* 1999; **38** :4389–4401.

[20] Gordon C. A generalization of the maximum noise fraction transform. *IEEE T. Geosci. Remote* 2000; **38** :608–610.

[21] R. Bhargava “Practical FTIR chemical imaging for cancer pathology” *Anal. Bioanal. Chem.*, **389**, 1155-1169 (2007)

Anal Bioanal Chem (2007) 389:1155–1169
DOI 10.1007/s00216-007-1511-9

ORIGINAL PAPER

Towards a practical Fourier transform infrared chemical imaging protocol for cancer histopathology

Rohit Bhargava

Received: 30 May 2007 / Revised: 9 July 2007 / Accepted: 12 July 2007 / Published online: 5 September 2007

Abstract Fourier transform infrared (FTIR) chemical imaging is a strongly emerging technology that is being increasingly applied to examine tissues in a high-throughput manner. The resulting data quality and quantity have permitted several groups to provide evidence for applicability to cancer pathology. It is critical to understand, however, that an integrated approach with optimal data acquisition, classification, and validation is necessary to realize practical protocols that can be translated to the clinic. Here, we first review the development of technology relevant to clinical translation of FTIR imaging for cancer pathology. The role of each component in this approach is discussed separately by quantitative analysis of the effects of changing parameters on the classification results. We focus on the histology of prostate tissue to illustrate factors in developing a practical protocol for automated histopathology. Next, we demonstrate how these protocols can be used to analyze the effect of experimental parameters on prediction accuracy by analyzing the effects of varying spatial resolution, spectral resolution, and signal to noise ratio. Classification accuracy is shown to depend on the signal to noise ratio of recorded data, while depending only weakly on spectral resolution.

Keywords Fourier transform infrared spectroscopy · FTIR imaging · Infrared microscopy · Prostate · Histopathology · Microspectroscopy

R. Bhargava (✉)
Department of Bioengineering and Beckman Institute for Advanced Science and Technology,
University of Illinois at Urbana-Champaign,
Urbana, IL 61801, USA
e-mail: rxb@uiuc.edu

Introduction

Cancer is one of the leading causes of death in the western world and is becoming increasingly prevalent worldwide. It is well established that appropriate therapy for cancers diagnosed early generally leads to improved prognosis and longer survival. Consequently, population screening tests to detect disease are increasingly being deployed. The emphasis in screening populations is on obtaining a high sensitivity through simple diagnostic tests. For example, the prostate-specific antigen (PSA) assay [1] helps triage persons at risk for prostate cancer. A cutoff level (typically 4 ng mL^{-1}) or increase in PSA velocity implies that the screened person should be at heightened surveillance and typically undergoes a biopsy to confirm disease. Morphologic structures in biopsied tissue, as diagnosed by a pathologist, are the only definitive indicator of disease and form the gold standard of diagnosis [2]. Along with clinical history, stage, and PSA values, pathologic diagnoses form a cornerstone of clinical therapy and serve as a basis for a vast majority of research activity [3].

Typically, multiple samples are withdrawn from the organ during biopsy. Extracted tissue samples are fixed, embedded, and sectioned (typically to 1- to 5-μm thickness) onto a glass slide for review. By itself, tissue does not have much useful contrast in optical brightfield microscopy. Hence, the prepared slide is stained with dyes. A mixture of hematoxylin and eosin (H&E) is commonly employed, staining protein-rich regions pink and nucleic acid-rich regions of the tissue blue, for example, as shown in Fig. 1. Using the contrast, a trained person can recognize specific cell types and alterations in local tissue morphology that are indicative of disease. In prostate tissue, epithelial cells line three-dimensional ducts. In two-dimensional thin sections, thus, the cells appear to line empty circular regions (lumen).

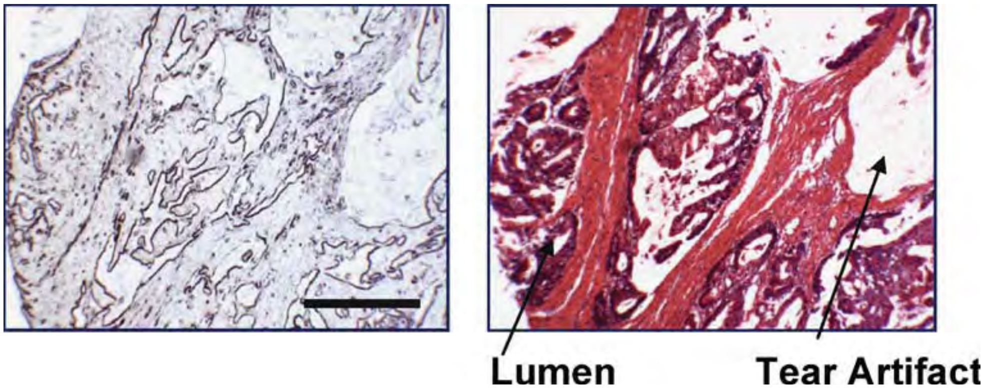


Fig. 1 Brightfield microscopy images of unstained (*left*) and stained (*right*) prostate tissue sections. Hematoxylin and eosin (H&E) stains provides contrast, allowing a trained person to recognize epithelial cells and ductal structure (lumen), while ignoring artifacts and confounding morphologies. A trained human can also learn to robustly recognize patterns within lumen that indicate cancer. The *scale bar* corresponds to 100 μm

Distortions in normal lumen appearance provide evidence of cancer and characterize its severity (grade). The process is fundamentally a manual pattern recognition that seeks to match observations to known healthy or diseased morphologies.

Manual examination of biopsies is very powerful in that humans can not only recognize disease generally but can also overcome confounding preparation artifacts, detect unusual cases, and recognize deficiencies in diagnostic quality This capability of considering and neglecting features based on prior knowledge is crucial for accurate and robust diagnoses. The process, however, is time consuming, allows for limited throughput and, frequently, leads to variance in subjective judgments about the disease severity, i.e., grade [4]. As an alternative, computer-based pattern recognition approaches to diagnose disease may provide more accurate, reproducible, and automated approaches that could reduce variance in diagnosis while proving economically favorable. Hence, attempts have been made to characterize morphology using H&E image analysis as well as biomarkers to stain for specific molecular features. Automated approaches that can rival human performance in usual clinical settings, however, are still unavailable. Specifically, the attributes of high accuracy and robust applicability are lacking.

The information content of H&E-stained images is limited and attempts to automatically recognize structural patterns indicative of prostate cancer, unfortunately, have not led to clinical protocols. Similarly, probe-based molecular imaging can provide exquisite information regarding the location and content of specific epitopes but is limited by complex diseases not expressing universally the same epitopes or panels of markers. Stains used can generally detect one feature that may aid diagnosis (e.g., AMACR) but do not provide entire diagnostic information in themselves. An exciting alternative is emerging in the form of chemical imaging and microscopy [5]. As opposed to conventional dye-assisted imaging or probe-assisted molecular imaging, chemical imaging [6] seeks to directly measure the identity and/or concentration of chemical species in the sample using spectroscopy. Hence, no molecular probes (MPs) are needed to see the presence of specific epitopes; instead computer algorithms are used to extract information from the data (instead of MP hybridization) and statistical methods are used to provide confidence (as opposed to brown tints for MPs). The approach is limited only by the ability of the technology to sense specific types of molecules or otherwise resolve chemical species and morphologic structures. Among the prominent approaches are vibrational spectroscopic imaging, both Raman and infrared (IR), as well as mass spectroscopic imaging (MSI) [7, 8] and magnetic resonance spectroscopic imaging (MRSI) [9]. While each technology promises a specific measurement (e.g., proteins or metabolic products) for specific situations (e.g., in vivo or ex vivo), IR spectroscopic imaging [10] is particularly attractive for the analysis of tissue biopsies in that it permits a rapid and simultaneous fingerprinting of inherent biologic content, extraneous materials, and metabolic state [11–14].

IR spectroscopic imaging, generally practiced using interferometry and termed Fourier transform infrared (FTIR) spectroscopic imaging or, succinctly, FTIR imaging, offers a particular combination of spatial, spectral, and chemical detail [15]. Limitations of FTIR imaging include coarser spatial resolution compared to Raman imaging or high powered optical microscopy and lack of specific molecular detail compared to MSI. Tissue biopsies are examined as thin sections on a solid substrate. The tissue is dehydrated and is stable due to fixation. Typically, structures of pathologic interest are several to hundreds of micrometers in size, requiring fairly moderate magnifications for decision making. These conditions imply that the need to image in vivo, at exceptionally high spatial resolution, or in aqueous environments is not critical and that standard pathologic laboratory processing can be employed for IR imaging. Due to the linear absorption process being utilized, the signal from IR spectroscopy is large and readily obtained, promising relatively simple instrumentation. Hence, the technology provides a platform that is potentially useful for clinical practice in pathology. It must be emphasized that no particular technology is ideally suited to all applications but a careful matching of the

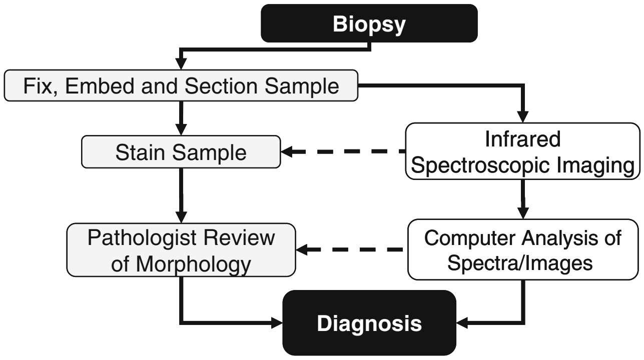

Biopsy
Fix, Embed and Section Sample
Stain Sample
Infrared Spectroscopic Imaging
Pathologist Review of Morphology
Computer Analysis of Spectra/Images
Diagnosis

Fig. 2 Potential application of FTIR imaging for pathology. The current paradigm of cancer diagnosis and grading upon biopsy involves sample processing, staining, and pathologist review (*left*, *shaded boxes*). To implement the paradigm of automated analysis (*right*, *unshaded boxes*), IR chemical imaging is followed by computer analysis for diagnosis. Since IR imaging is label-free and non-perturbing, the sample can be stained, providing the pathologist with both IR chemical and conventional stained images

technique to the application can lead to useful protocols. While the potential advantages of FTIR imaging for examining tissue biopsies is high, practical protocols for clinical deployment are being developed by many groups.

Numerous recent reviews are available to address biomedical applications of FTIR spectroscopy and imaging [16–20], especially related to diseases and cancer. These reviews address instrumentation, the applicability to various systems, spectroscopic bases and classification algorithms for decision making, and controversial aspects in the backdrop of the evolution of the field. The commercial availability of high-fidelity FTIR imaging instruments, advances in computers and data analysis algorithms, and increasing interest have combined to generate an increasing volume of studies. At the same time, there is considerable debate emerging on various aspects of the process. Reports study a variety of organs that may not correlate in behavior, utilize different sample acquisition and processing techniques, employ different instrumentation, data acquisition, or handling protocols, and apply a variety of decision-making algorithms. While this has led to a lively community of practitioners and exploration of various facets such as resolution, biological diversity, and chemometric or statistical methods, studies have generally focused on one aspect. Many excellent studies have developed each of these aspects to the point of routine use in advanced laboratory. The focus in the field is now on understanding biochemical signals and developing protocols from high quality data that can actually lead to clinical acceptance. We contend that the development of clinical protocols is necessarily integrative and, in this manuscript, review first the salient aspects in developing a practical, integrative approach to spectroscopic imaging for cancer histopathology. Second, we discuss the issues of spatial selectivity, sample size calculations,

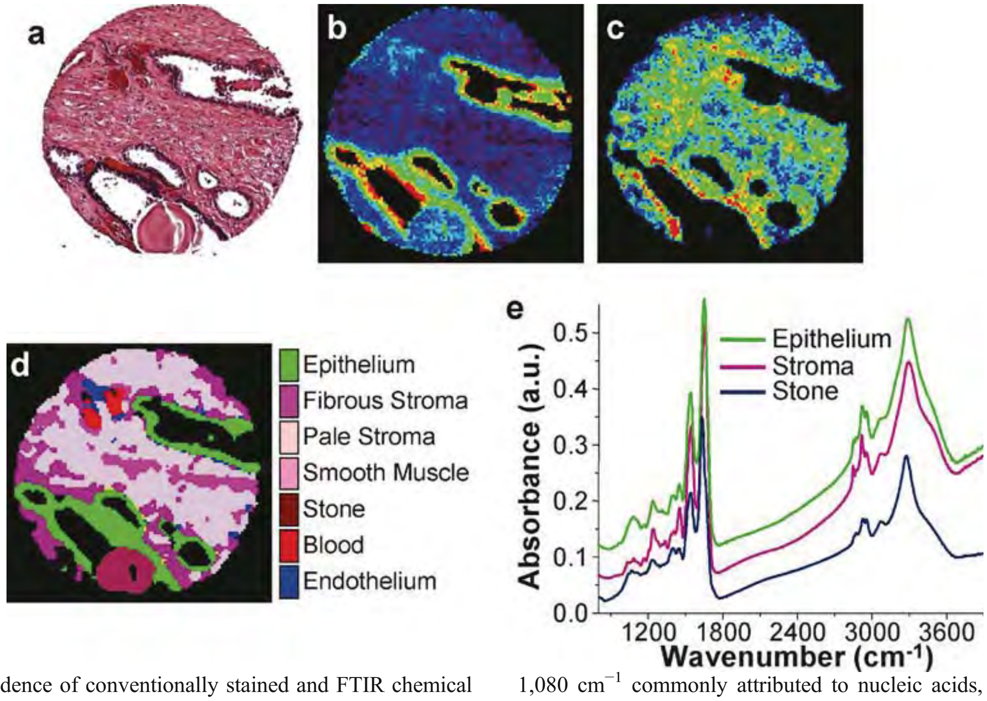


Fig. 3 Correspondence of conventionally stained and FTIR chemical images for pathology applications. **a** Hematoxylin and eosin (H&E)-stained image of prostate tissue section. Hematoxylin stains negatively charged nucleic acids (nuclei & ribosomes) *blue*, while eosin stains protein-rich regions *pink*. The diameter of the sample is ca. 500 μm. Simple univariate plots of specific vibrational modes provides for enhancement or suppression of specific cell types. **b** Absorption at 1,080 cm^{-1} commonly attributed to nucleic acids, highlights nuclei-rich epithelial cells in the manner of hematoxylin. **c** Spatial distribution of a protein-specific peak (ca. 1,245 cm^{-1}) highlights differences in the manner of eosin. The entire spectrum can be analyzed for a series of markers that provide more information than H&E or univariate images, as shown in **d** where specific cells are color coded based on their spectral features (**e**)

optimization considerations, and potential improvements in algorithms that can provide faster results. Tests to determine performance and limits of accuracy are reported as a function of experimental parameters. We focus here on prostate histology as an illustrative test case, but emphasize that the approach is applicable and similar insight is gained with other tissues [21]. Further, exciting results have recently been reported for diagnosis, grading, and classification of prostate cancer [22–26], including the effects of zonal anatomy [27] and cytokinetic activity on spectra [28]. An extension of the methodology here to pathology will help formulate better protocols and allow a better understanding of the performance of classifiers.

Approach and essentials

The promise of chemical imaging for pathology is illustrated in Fig. 2. Our approach has been to attempt integration of our developments with current clinical practice. Hence, we employ tissues that have been biopsied, fixed, embedded, and sectioned as per usual clinical protocols. We differ in the de-paraffinization step, suggesting a gentle wash with hexane and do not stain the tissue. Additionally, as IR chemical imaging only employs benign light, it is non-perturbing and entirely compatible with all downstream pathology processes. Hence, the sample may be stained as usual (Fig. 2, dashed arrow, top). Visualizations similar to those observed in conventional pathology are possible without staining the tissue. For example, Fig. 3 correlates H&E and infrared spectral images. Visualizations similar to H&E images may be “dialed-in” by utilizing specific spectral features indicative of tissue chemistry. Although, the IR data only demonstrate univariate representations in the images, automated mathematical algorithms can determine the cell types and their locations within the image, while providing quantitative measures of accuracy and statistical confidence in results [29]. These data may be employed to directly provide diagnoses or to inform the pathologist (Fig. 2, dashed arrow, bottom), helping them make better decisions. Since the results are images, information exchange between spectroscopists and clinicians is facilitated. Spectroscopic analyses can potentially be fully automated; thus, no additional users need to be trained or knowledge base acquired by current clinicians.

A major challenge in the field is the development of robust algorithms that employ spectral data to provide histopathologic information. Both supervised and unsupervised approaches have been employed. We believe that unsupervised methods are more suited to research and discovery. Supervised methods are preferred when the data need to be related to known conditions, e.g., clinical diagnoses. The development of supervised classification of IR chemical imaging data for histopathology is fairly straightforward [30]. The process is shown in Fig. 4. First, a model for classification is selected. The model comprises all possible outcomes for any pixel in the images and is, hence, bounded by definition. We term each histologic constituent of the model a class to denote that it may not correspond to specific cell types or entities corresponding

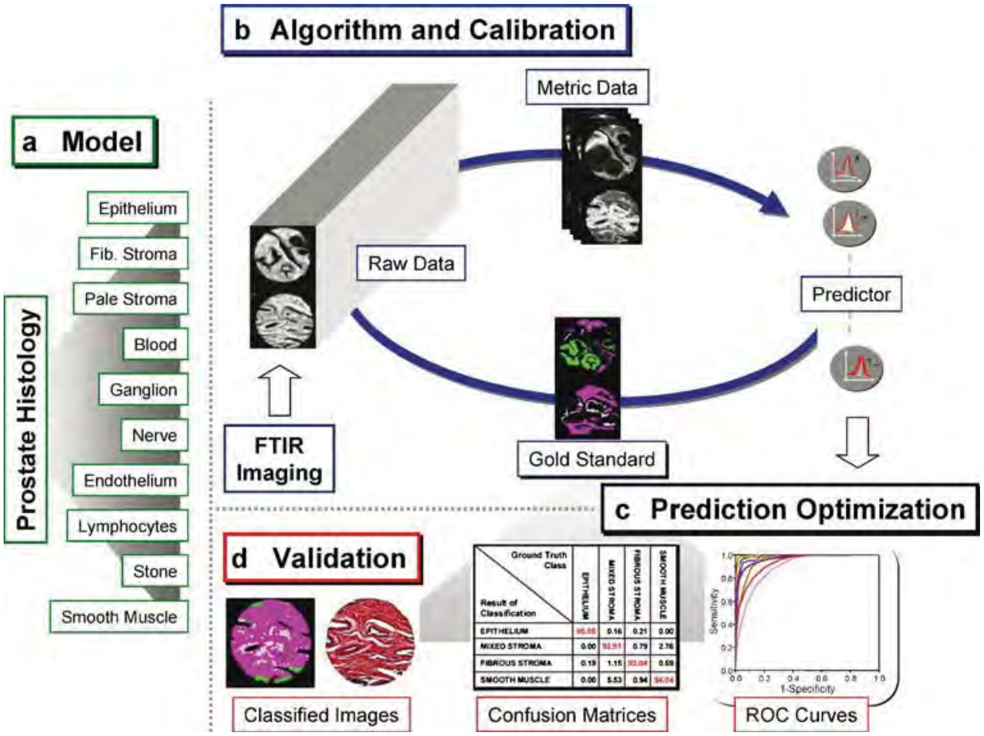


Fig. 4 Process for relating pathologic or physiologic state to FTIR chemical imaging data. A model is chosen for supervised classification (**a**). **b**–**d** Training data is reduced in size and optimized into a prediction algorithm using gold standard data. The developed algorithm is validated against a second, independent data set and the accuracy is measured using three different methods: ROC curves, confusion matrices, and image comparisons

to morphology-based pathology. While this allows for simplifications and allows the user to focus on specific cells relevant in disease, it is also likely to prove useful in the discovery of different chemical entities that appear morphologically identical.

Next, data from a large number of tissue samples is recorded. A set of pixels are specifically marked (gold standard) by different colors to correspond to known regions of tissue, usually by comparison with an H&E-stained image or with immunohistochemically stained images [21]. The recorded data set is reduced to a smaller set of measures that capture the classification capability of the entire data set. We termed these measures metrics. There are numerous means of obtaining the metric data set: manual selection of large spectral regions, principal components analysis, genetic algorithms, or a sequential forward selection algorithm. A numerical algorithm is then chosen, for example, a linear discriminant analysis, neural network, SIMCA, or modified Bayesian classifier [31]. The classifier is optimized iteratively, if needed, to optimally predict the training data set. Subsequently, the algorithm is applied to a second data set (independent validation) that has been independently marked for each class. A comparison of the gold standard marking with the computationally predicted class provides a measure of the accuracy. We have employed three measures of accuracy: receiver operating characteristic (ROC) curves [32] that represent the sensitivity and specificity trade-off of the classifier, confusion matrices that provide the fraction of pixels of each class classified as pixels of all classes, and classified images that can be compared pixel-for-pixel to other images. Additionally, it is often instructive to drill into the classifier to obtain the basis for classification or the distribution of confidence intervals for various samples. The last two factors are generally not apparent in previous studies.

There are three key developments that are needed for this approach to be successful: (a) high-fidelity FTIR imaging instrumentation, (b) high-throughput sampling, and (c) robust classification that provides statistically significant results in a manner that can be appreciated by non-experts in spectroscopy. We briefly review the three developments next.

FTIR imaging

Need for spatially resolved data

The need for spatially resolved data has been recognized [33], but the effect of limited resolution data on classification is not entirely clear. The primary complication of coarse spatial resolution, obviously, arises from boundary pixels. These can be defined as pixels that are assigned to one class but would likely yield more classes, to their physical limits, were finer resolution available. As a consequence, the spectral content of the boundary pixel is likely to be mixed and will likely lead to errors in classification. For example, the confounding contribution of stromal spectra to cancerous epithelial cells in breast tissue has been proposed [34]. As the resolution becomes coarser, the fraction of pixels in an image that belong to boundary pixels increases. Inclusion of these pixels has been shown to be a primary contributor to error rates in data [29], while their exclusion in accounting for accuracy necessarily implies that not all pixels are included. We sought to examine the effect of spatial resolution on the prevalence of boundary pixels.

We binned data acquired at 6.25-μm pixel size from 148 samples in a validation data set (≈7000 pixels/sample) to 10-, 15-, 20-, 30-, and 50-μm pixel sizes. There is an important distinction between pixel size and spatial resolution. The pixel size denotes the best possible optical resolution, which may be limited by longer wavelengths in the spectrum and optical effects to yield a poorer measured resolution [35–38]. For each dataset, we classified the tissue and determined neighbors of each pixel that did not belong to the class of the pixel. Some pixels that have no neighbors of other classes may still have empty pixels as neighbors. Since neighboring empty pixels can only provide optical distortion [39] but do not affect spectral content; we do not consider them further. The number of neighbors for epithelial pixels for different spatial resolutions may be seen in Fig. 5. The first observation is that a large majority of pixels have the same class pixels as all eight neighbors. The fraction of pixels with all neighbors of the same class

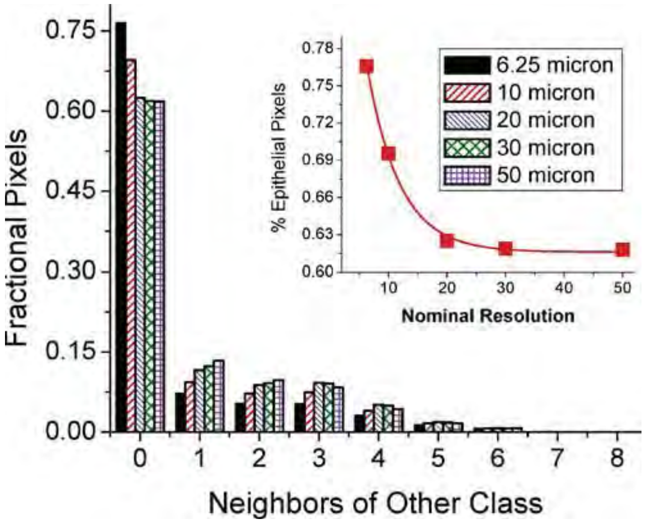


Fig. 5 Neighbors of cell types other than epithelium or empty space for different spatial resolutions. The *inset* shows the decrease in percent epithelial pixels that do not have any other cell types as neighbors

decreases rapidly with decreasing resolution and stabilizes at ca. 20 µm. Hence, a spatial resolution coarser than 20 µm is unlikely to have an effect on the classification but is expected to lead to about 25% more epithelial pixels being contaminated compared to 6.25-µm pixel sizes. The precise effect on a specific sample is very dependent on the sample morphology and is generally associated weakly with pathologic state. While in itself, the statistic does not imply that results from coarser resolution studies will be invalid, practitioners must recognize that error rates may be higher and that this contribution may be mitigated by using commonly available imaging systems.

One danger of classifying mixed composition pixels is whether they may be classified as an entirely different class or disregarded from the data set as belonging to no class. We simulated pixels of composition ranging from 0 to 100% for pairs of each class. We also added noise to simulate different data acquisition conditions. An example of the data can be seen in Fig. 6. Average spectra, one each from the two classes, are baselined and added in ratios varying linearly from 0 to 100%. Figure 6b demonstrates the classification of the gradient data set. In general, the classification works well, favoring the class with higher concentration. The classifier is also stable at the noise levels examined. A surprising result is that pixels between epithelium and fibroblast-rich stroma are classified as mixed stroma. This drawback, however, is the only example of two classes mixing to yield an entirely different one. The reason also stems from the definition of the mixed stroma class. While the class was designed to handle those stromal cells that were not clearly fibroblasts or smooth muscle in origin but appeared mixed, a mix of epithelium and fibroblast-type stroma also leads to the classification as mixed stroma. Noise seems to have little effect on this behavior.

The full simulation of all classes (not shown) reveals that mixed pixels generally can be classified as the constituent classes with the higher concentration. Clearly, boundary pixels at epithelial fibroblast-rich regions must be handled with care. The increase in boundary pixels at lower resolution also implies that this type of systematic misassignment may arise more frequently. The rate of occurrence of boundary pixels may be even lower for synchrotron-based imaging that is conducted at higher pixel density or in emerging approaches that utilize synchrotron-based interferometers and array detectors. The simulated example above, however, demonstrates that simply oversampling a spatial region to increase pixel density may allow for better definition of the interface and assignment of pixels, though it will not address spectral purity. Hence, for analyses based on spectral discrimination, mixture models will have to be developed based on entire spectra. For example, multivariate curve resolution techniques hold promise.

A further complication arises in using data from histologic classification for pathologic diagnoses. For example, the boundary epithelial pixels classified above may disproportionately contribute to classification errors. We have found evidence for the same in studies for both cancer pathology and for histology in tissue from different organs. For example, the boundary pixels in benign tissue get

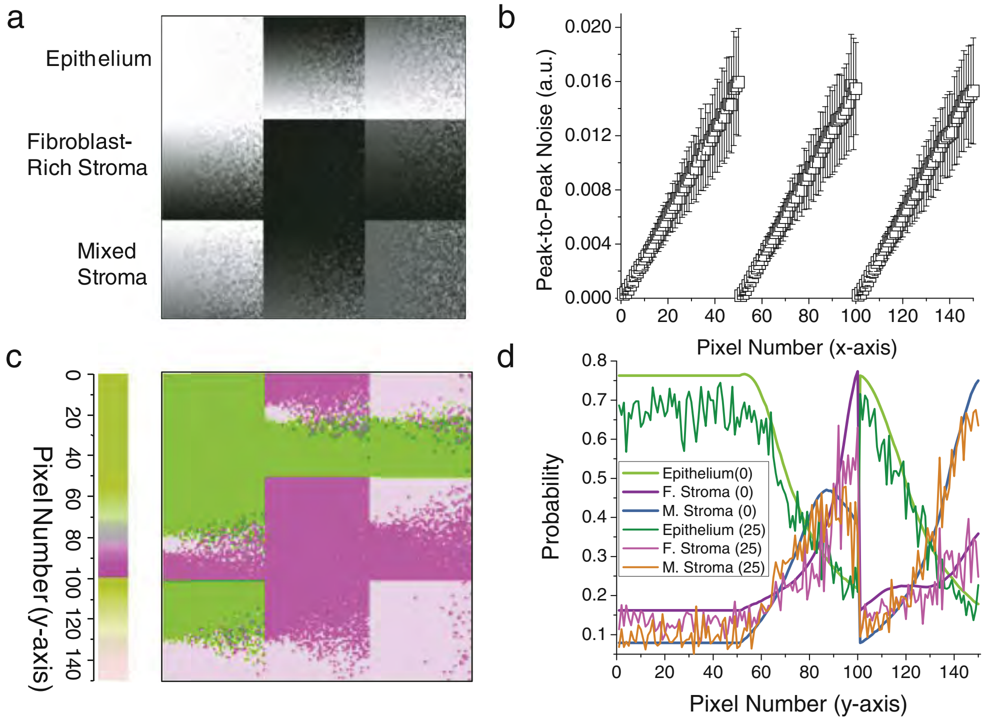


Fig. 6 Mixture models and classification for prostate histology. **a** Absorbance at 1,080 cm^{-1} for three classes and their mixtures. The *first column* contains mixtures of epithelial cell spectra with the average spectrum from fibroblast-rich stroma and mixed stroma. The *second* and *third columns* contain mixtures with fibroblast-rich and mixed stroma, respectively. The concentration changes from 0 to 100% linearly along the *y-direction* as indicated by the *color bar* in **c**. **b** Along the *x-axis* of the composite image, the noise in each cell increases linearly. *Error bars* are standard deviations of noise in the spectra. **c** Classified image for the data, demonstrating the effect of composition and noise on classification. **d** Probability profiles of the three cell types at columns 1 and 25, demonstrating the effect of noise

Table 1 Correlation of composition for samples between 6.25-mm pixel sizes and other pixel sizes

Pixel size (micron)	Epithelium	Fibroblast-rich stroma	Mixed stroma
10	0.9913 x (0.9976)	0.9847 x (0.9923)	1.0300 x (0.9957)
20	1.0156 x (0.9906)	0.9671 x (0.9775)	1.0473 x (0.9787)
25	1.0404 x (0.9896)	0.9768 x (0.9624)	1.0262 x (0.9617)
30	1.0720 x (0.9773)	0.9683 x (0.9507)	1.0175 x (0.9363)
50	1.1180 x (0.9459)	0.9410 x (0.8947)	1.0390 x (0.8723)

The first row in each cell denotes the composition factor for that pixel size and class. For example, for every 100 μm^2, the area of epithelial pixels at 10-μm pixel size is 99.13% of that at 6.25-μm pixel size. Increasing/decreasing numbers represent pixels being increasingly/decreasingly classified as that class. The ratios are not uniform for every sample and the regression coefficient of the best fit line passing through the origin is provided in the second row of the each table cell. Increasing pixel sizes reflect greater variance from the fit line

misclassified as cancerous, leading to the major source of error in applying this approach to pathology. At this time, the evidence is anecdotal and needs further investigation to quantify the extent of the error and its mitigation by advanced numerical processing. The last interesting aspect of lower spatial resolution is that it tends to over-predict certain classes. For example, Table 1 demonstrates the regression results of each samples composition against that obtained at 6.25 μm for three classes. While the regression coefficient is high, it is clear that epithelial and mixed stroma fractions are overestimated and fibroblast-rich stroma is underestimated with decreasing pixel size. There are differences based on underlying pathology. For example, normal epithelium is generally encountered in 10- to 40-μm-wide strips, while high grade tumor may be hundreds of micrometers to millimeters in size. Individual sample variability reflected in the regression coefficient decreases with increasing pixel size. In spectroscopic models to predict diseases that include morphological units but are based on average spectra, mixed pixels may lead to estimates with large errors. For example, a 1:1 mixed region of epithelial and fibroblast pixels at 6.25-μm pixel size increases to ca. 1.19:1 for 50-μm pixel size. Hence, the use of histologic mixture models at limited spatial resolution may not be estimated correctly, providing evidence that the percentage content of cell types in a limited field of view is likely to be a less robust measure of tissue histopathology.

Evolution and capabilities of current instrumentation

To overcome confounding by mixing, as discussed above, microscopectroscopy was proposed as an alternative [40]. Single spectra (non-FTIR) have been recorded from microscopic samples for over 50 years [41] by restricting light incident on the sample through an aperture. More than one point, however, is required for tissue analysis. Hence, sequentially rastering the point at which spectra are recorded, termed mapping or point microscopy, was proposed [42]. A practical instrument obtained by coupling an interferometer, a microscope, and automated stage in the late 1980s [43] helped in numerous materials science [44], forensic [45], and biomedical [46, 47] studies. Unfortunately, the mapping approach has a number of drawbacks in realizing the goal of an FTIR microscopy analog to optical microscopy [48].

More than 85% of cancer arises in epithelial cells, which often form surface layers that are 10- to 100-μm wide. As we demonstrated in the previous section, however, a resolution higher than ca. 10×10 μm is preferable. Consequently, the illuminated spot at the sample has to be made smaller, throughput decreases proportionally, which in turn decreases the signal to noise ratio (SNR) of acquired spectra. Orders of magnitude brighter sources, e.g., synchrotrons, may be employed to recover the lost SNR. Unfortunately, synchrotron or free electron lasers [49] are prohibitively expensive and no laboratory lasers exist for the wide spectral region. An alternative is to average successive measurements (co-adding) to increase statistically the SNR. Since the SNR increases only as the square root of the number of averaged spectra, long averaging periods are required. The situation may be mitigated by using higher condensing optics, sources at higher temperatures, slightly faster scanning than used here,[1] gain ranging [50], or ultra-sensitive detectors [51]. Even if a hypothetical instrument with all these advances were constructed, ca. 10- to 20-fold reduction in time would be obtained. Furthermore, this calculation underestimates the time required by not considering losses due to diffraction or stage movement.

In prostate tissue, for example, the situation is similar to Fig. 1. Epithelial cells form 10- to 35-μm-wide foci around the cross-sections of ducts. Ducts appear as white circles in Fig. 1b, surrounded by epithelial cells that are depicted in blue. To analyze this morphology, aperture dimensions of ca. 6 μm×6 μm (≈ cell size) are proposed [31]; for this case, the mapping approach would require ca. 1,028 h for a

[1] There is no advantage to faster scanning once the modulation frequency has reached optimum level for MCT detectors (1 MHz). The reduced time to observe signal then decreases the SNR.

500 μm×500 μm sample [31]. Hence, mapping is not a viable option. In contrast to point mapping using apertures, large fields of view are measured in FTIR imaging. Contributions from different sample areas in imaging are separated by an array of mid-IR-sensitive detection elements in the manner of imaging with CCD devices for optical microscopy. By coupling the multichannel detection of focal plane array (FPA) detectors with the spectral multiplexing advantage of interferometry, an entire sample field of view is spectroscopically imaged in a single interferometer scan [52]. Depending on the microscopy configuration, thousands of moderate resolution spectra can be acquired at near-diffraction-limited spatial resolution in minutes [53, 54]. The time advantage over mapping is nominally the number of pixels in the FPA (16- to 65,000-fold) but the noise characteristics of FPAs are poorer than sensitive single point detectors [55]. Hence, the SNR-normalized advantage is lower [56]. Faster detectors are being used for imaging and promise significantly higher SNR in the same time. For example, we have employed a 128×128 element MCT array operating at ca. 16 kHz to acquire a full data set in ca. 0.07 s [unpublished]. These rates of data acquisition are approximately a factor of 10 higher than commercially available, but are required for practical data acquisition times. Increase in data acquisition speed remains a bottleneck for applications of IR imaging to routine clinical studies. Coupled with the complexity and cost of instrumentation, present technology provides preliminary capability but is likely to prove a barrier to practical clinical translation.

High-throughput sampling and statistical pitfalls

Quantitative analyses of results

The best imaging instruments (which employ sensitive detectors and a small multichannel advantage) can acquire data in about 0.1% of the time required for mapping for equivalent parameters. Hence, point mapping studies in pathology typically exceed numbers in only one of these categories: spatial resolution (ca. 15–20 μm), numbers of patients (ca. 50) or recorded small numbers of spectra per patient (ca. 100). These numbers may typically be improved an order of magnitude with imaging. For example, a recent report analyzed ca. ten million spectra from ca. 1,000 samples at a spatial resolution of 6.25 μm [26]. This quantitative validation is necessary for any automated biomarker approach (vide infra) [57]. Studies are underway in our and other laboratories to correlate spectral patterns with other physiologic and pathologic conditions; recent published studies verify the robustness and potentially wide applicability of FTIR microscopy [58, 59].

Sample size

Though these studies demonstrate potential, [60, 61] considerable debate exists on reproducibility and accuracy measures for larger studies [29]. The first response of many practitioners to new data is a question of validity based in limited statistical confidence. A detailed understanding is emerging from the work of several groups regarding appropriate sample control [62] and confounding factors due to biology [63]. Inherent differences between patient cohorts, effects of sample preparations and measurement noise are topics that can be addressed with the available imaging technology but are yet to be fully explored. Hence, validating robust spectral markers for large sample populations [64, 65] is exceptionally challenging and the chance for chance and bias influencing results exists.

Most importantly, the fundamental question of sample size required has remained open. There are two major concerns: first, the optimal sample size in forming calibration sets and a prediction algorithm. Second, investigators must determine whether the results shown can be supported by statistical considerations. While the first problem is essentially one of optimizing a model and prediction algorithm, the second impacts the quality of results and claims of applicability directly. In this manuscript, we examine only the second aspect. Determining the optimal sample size to form robust models is a more involved problem and is discussed elsewhere. The statistical validity of obtained results and dependence on data acquisition parameters are discussed later in this manuscript. Specifically, we estimate sample size based on the standard error for the area under the curve for an ROC curve.

Gold standard

The selection of pixels as gold standards needs great care. It must be done independently of any classifier training or validation, thus ensuring a blinded study design. Once the gold standard set is determined, it must not be changed. This will ensure that there is no bias in the process. Care must be taken to avoid pixels that do not lie on the tissue or those that are at the boundary as these may artificially inflate the error. The use of all pixels in an image has been suggested and their exclusion has been proposed to contribute selection bias. Selection bias, however, does not arise in pixels that are chosen independent of validation algorithms. The exclusion of boundary pixels is necessary in both training (to avoid spurious probability distribution functions) and validation (to prevent introduction of errors). There are major technological difficulties in relating stained visible to IR images from unstained tissue due to changes during staining, leading to errors. Hence, it has been proposed that the exclusion of boundary pixels in akin to the performance of a classifier with a reject option for the boundary.

Sampling, archiving, and consistency

While it is unclear what an optimal sample size would be, it is clear that a large number of tissue samples are needed for effective validation. While it may theoretically be possible to train on a single sample, validation of a protocol is required on more samples. We recognized that one does not need to observe the full surgically resected tumor for validating IR protocols, but would need a representative small section. Hence, we employed tissue microarrays (TMAs) [66] as a platform for high-throughput sampling. TMAs consist of a large number of small tissue samples arranged in a grid and deposited on the same substrate. They are typically manufactured by embedding cylindrical cores in a receiving block and sectioning the block perpendicular to the long axis of the core. Thin sections are then floated on to a rigid substrate for analysis. The technique facilitates rapid visualization of results of any classification protocol, while revealing localization and prevalence of any errors. Sample processing times may easily be increased 100-fold, valuable tissues are optimally utilized, and consecutive TMA sections can be used to correlate with staining results. Construction and analysis of TMAs has been automated, further increasing the throughput. For spectroscopists, TMAs provide a ready source of tissue to test hypothesis and develop prediction models.

The validity of employing TMAs for prostate cancer research and, especially, for cancer grading has been addressed by a number of authors [67]. For example, a study of genitourinary pathologists [4] with images from TMA cores assesses that ca. 90% considered this approach useful for resident training and for pathology teaching. Further, Gleason score was easily assigned to each TMA spot of a 0.6-mm-diameter prostate cancer sample. Hence, the utility of TMAs is not only in providing numerous samples in a compact manner for the advantages above, but also in consistency of the diagnoses and precision in analyzing similar areas. Virtual tissue microarrays could be constructed from different areas of large samples, thus providing many sub-samples for within-patient and among-patient comparisons. This approach has not yet been reported but is likely a useful extension of the TMA concept.

Prediction algorithms and high-throughput data analysis

Univariate algorithms

The major technological advances of fast FTIR microscopy and high-throughput tissue sampling have been addressed by imaging and TMAs respectively. There is still some confusion and widespread disagreement, however, about the “best” approach to extract histopathologic information from FTIR imaging data. Several early manuscripts employ univariate correlations to disease states [68]. While the results were exciting, it is now realized that they were statistically flawed and did not necessarily contain a fundamental basis in cancer biology. To our knowledge, there is no manuscript that has expressly demonstrated, using statistics arguments, why univariate analyses are likely to fail. There is widespread consensus and anecdotal evidence, however, among practitioners that argues against the approach. Consider the distributions for a univariate measure (absorbance at 1,080 cm^{-1} that is normalized to the amide I peak height) for benign and malignant cases as shown in Fig. 7.

The normalized histograms reveal that for specific, single samples the distribution of absorbance at pixels is such that it clearly indicates the metric to be a good one for cancer discrimination. When the distribution from all samples is considered, however, there is little difference in the distributions. Hence, many univariate measures described in the literature do not hold up in wide population testing. A TMA-based, high-throughput validation can easily prove that the measure is not a good one but does discriminate some samples. In Fig. 7, a cutoff value can generally be found that distinguishes disease, leading to the erroneous conclusion that the feature is universally indicative of disease state. Since a typical infrared spectrum has numerous frequencies and even non-chemically specific features that can provide discrimination, a small number of samples increases the probability of finding such discrimination by chance alone. Univariate measures that appar-

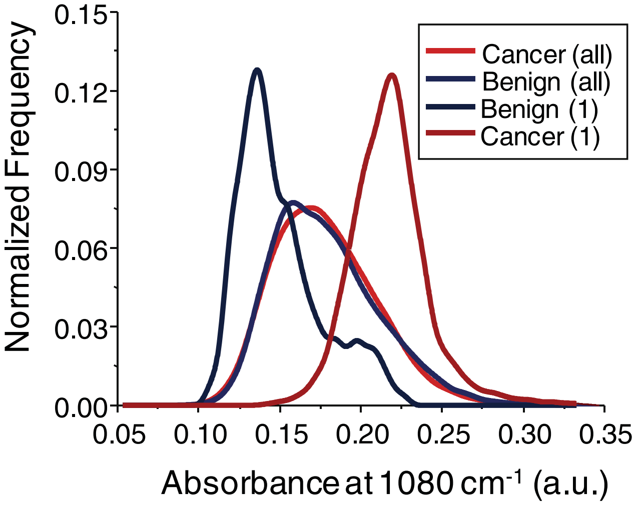


Fig. 7 Distribution of absorbance for individual spots and all pixels from each class, normalized by the total number of pixels in the class, demonstrates that the examination at patient level and at a global level may not correspond

ently provide discrimination when none exists can be equated to the false discovery rate (FDR) [69] of metrics. The FDR is very different from the p-value for determining that a metric separates two distributions; a much higher FDR can be tolerated than can a p-value. Similarly, a false negative rate has been proposed [70], which is not critical for our case as we have observed high accuracy without use of any erroneously left out metrics. While detailed calculations and their underlying concepts are too lengthy to reproduce here, for the sake of completeness, it suffices to say that for the expected number of metrics demonstrating discrimination, the FDR tends to zero for larger than ca. 30 samples. While correlations due to chance can be minimized by this approach, there is potential for unknown bias or error in prediction for small numbers of samples. Hence the algorithm must be integrated with sampling considerations.

Multivariate algorithms

It was argued in the previous section that univariate analysis may not provide a good measure of the population distribution. It can alternatively be argued that the individual differences in univariate measures are masked if population measures of the same are employed. Similarly, multivariate techniques may mask the individual measures in population testing. Hence, our philosophy has been to employ a multivariate, supervised classification in which the metrics are derived from univariate analyses. This enables us to carefully examine each metric for both population as well as individual sample relevance. While unsupervised clustering approaches provide good insight into spectral similarity, a supervised method forces a relation to common clinical knowledge. For example, as shown in Fig. 4 for prostate tissue, we consider a ten-class model to determine histology. The drawback is that the sensitivity of the approach to individual samples is lost at the expense of generality. One could potentially combine clustering and supervised classification. Clustering information on the training data set would emphasize individual sample distributions, which would allow for supervised classification tailored to each cluster type. Such an approach has not been implemented yet but is being attempted in our laboratories to classify samples optimally.

Dimensionality reduction

It is well recognized that the spectrum at each pixel needs to be reduced to a smaller set of useful descriptors that capture the essential information inherent in the spectrum. The reduction of full spectral information to essential measures helps eliminate from consideration those spectral features that have no information (non-absorbing frequencies), little biochemical significance (e.g., apparent absorption at non-chemically specific frequencies), inconsistent measures that may degrade classification, and those with redundant information. The number of useful measures is significantly smaller than the number spectral resolution elements and, hence, the process is also termed dimensionality reduction. Dimensionality reduction and further refinement (vide infra) also helps reduce the incidence of prediction by chance alone, reduce computation time and storage requirements. Potential measures of a spectrum's useful features are termed metrics and are defined manually in our scheme.

It may be argued that the metrics are not selected in an objective manner due to a human performing this task and some computer routines must be employed. While the use of an automated computer program is most certainly objective and reproducible, the algorithm that drives such programs is generated from spectroscopy knowledge. A well-trained spectroscopist can recognize spectral features and assign them to appropriate their biochemical basis. While a computer algorithm may be able to enhance subtle features in the spectrum, automated peak-picking algorithms run the risk of substantial error as they are based on some very specific criteria that may not be universally valid. We believe that computer algorithms are more suited to finding correlations and patterns that a human cannot for the sheer size and complexity of data. Hence, the process of determining which spectral features to consider is entirely manual in our approach. It must be emphasized that the universal set of metrics is selected manually but that the data reduction step to a set of metrics to be used in algorithms is entirely based on objective algorithms. Manual refinement of metrics for classification is, obviously, not recommended for possibilities of overlooking specific features, biasing the selection to specific feature sets, or in determining the optimal set of metrics for a classifier. Dimensionality reduction is also intimately linked to the data quality and classification algorithm employed.

Classification algorithm

A number of supervised algorithms have been applied to dimensionally reduced data, including those based on linear discriminant analysis, neural networks, decision trees, and modified Bayesian Classifiers. An intermediate step in some of these algorithms provides for a fuzzy result in which every pixel has a probability of belonging to every class. For example, in our approach, each pixel can have a probability (between zero and one) of belonging to each class. A discriminant function then assigns each pixel to a class based on a decision rule. The pre-discriminant data set, termed rule imaging set, contains important information. In our algorithm, it is a direct measure of the probability of the pixel belonging to the class. Hence, the probability value may be used to compare the potential of two protocols to distinguish a cell type or to quantify confidence in results for tissue classified by different methods.

Measures of accuracy and optimization

We prefer the use of the AUC for both optimizing algorithms and for validating results. Confidence in the value of the AUC, hence, is the primary test for the validity of developed algorithms and is characterized by the standard error of the value. For example, in validating the discrimination of epithelial from stromal pixels in a blinded validation set, the cumulative distribution of AUC in a TMA is shown in Fig. 8. More than 20% of the spots had an AUC >95% and no AUC value below 0.8 was recorded. One drawback of using ROC curves and AUC values is that the results are valid for one at a time classification. Hence, we have analyzed here the segmentation of epithelium from all other cell types. The tissue is classified into ten classes as before but the results are lumped into epithelial and non-epithelial pixels. Further, not all TMA cores have all types of cells. Hence, the two-class model also allows us to examine a large number of samples. Last, we excluded cores that did not contain at least 100 pixels of each class to leave 103 cores for the analysis.

Quantitative measures of performance and accuracy are perhaps the weakest portion of reports using IR spectroscopy for cancer pathology. Typically, sensitivity and specificity have been employed as summary measures. While these are indeed very relevant, we demonstrate that they are insufficient and classification analysis must utilize more measures to understand the process. Specifically, the use of receiver operating characteristic (ROC) curves [71] is an excellent direction. The area under the ROC curve is a further summary measure that provides both a quantitative understanding of the discrimination potential of the model and a convenient measure to compare multiple classification models. The third tool we introduced was the confusion matrix. While ROC curves provide the potential for correct classification of a binary rule at a time, confusion matrices correspond to a particular point on the ROC curve under the constraints of accuracy measures of other classes. These also directly correspond to the final segmentation of the rule image under an optimization condition. The optimization condition may simply be the maximization of the accuracy or may be the minimization of certain types of errors.

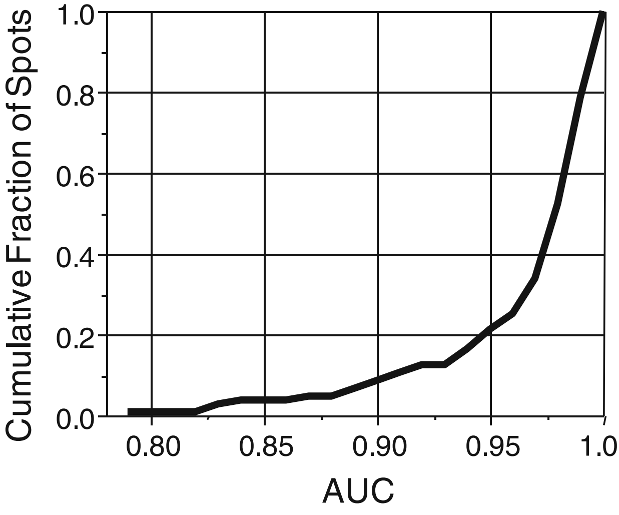


Fig. 8 Distribution of AUC values in a TMA for discriminating epithelium from stroma using the ten-class model

Discriminant and class assignment

In a multi-class analysis, our approach to evaluating ROC curves for a class is one at a time, i.e., all other classes are essentially lumped in the rule data and the highest probability of the lumped ensemble is compared to the class whose ROC curve is being built. Hence, the AUC values must be regarded as a potential for classification. They are best suited to answer the binary question of whether a pixel is correctly identified or not when considering a single class. This method is ideally suited to a cascaded classifier one at a time. Such a classifier has not been reported yet but would provide a means to explicitly determine the error for any given classification scheme.

Experimental parameters and classification

Here, we take advantage of the trading rules of FTIR spectroscopy and imaging to model the effects of the experimental parameters on the classification process. While the signal to noise ratio (SNR) and resolution are generally arbitrarily fixed in most studies, we demonstrate their importance in classification.

Effect of signal to noise ratio

There are two issues: what is the “best” SNR to formulate algorithms and second, provided an algorithm, what is the least SNR that would provide adequate classification. Only the latter issue is examined here. As with conventional FTIR spectrometers, imaging spectrometers obey the trading rules of IR spectroscopy. Hence, if an n-fold reduction in SNR provides the same results, data acquisition will be n^2-fold faster. Thus, in addition to an interesting fundamental behavior of the classifier, the role of SNR has a direct bearing on the speed at which data is acquired.

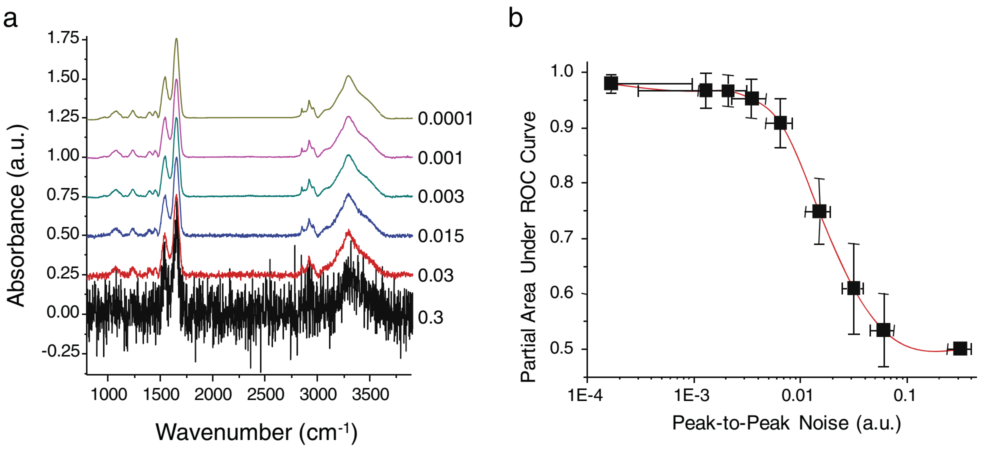


Fig. 9 **a** Noise in the data set as a function of added random noise. **b** Effect of spectral noise on the accuracy of classification as measured by AUC values

We examined classification accuracy as a function of average spectral noise. To strictly examine the effect of noise, data must be acquired at different co-added spectral numbers. The time required for imaging an array multiple times, however, is prohibitive. Hence, we computationally added random, Gaussian noise to the original spectral data. Peak-to-peak and root mean square (rms) noise were measured in the 1,950- to 2,150-cm^{-1} region adjacent to the amide I peak.[2] Representative single pixel spectra from the data sets are shown, as a function of noise, in Fig. 9a. We additionally plotted the observed noise levels against the added noise to verify linearity (plot not shown). The linear relationship conforms to the expected result and provides a scaling factor to express the equivalent reduction in data acquisition time (co-addition) that would be realized at that noise level. For example, the addition of 0.005 a.u. of noise raises the peak-to-peak noise from 0.0013 to 0.015 a.u., corresponding to a decrease in data acquisition time by a factor of ca. 100 for this data set. In addition to increasing noise, we employed an algorithm based on an MNF transform [72, 73] to mathematically eliminate noise. The observed peak-to-peak noise was 0.00017 a.u., corresponding to an increase in data acquisition time by a factor greater than ca. 100. Hence, the data examined span about 5 orders in magnitude of collection time.

The average height of the amide I peak was 0.42 a.u. in all cases, providing a SNR of 2,500 (MNF-corrected data) to 1.5 for the data sets. Accuracy as a function of the noise level is shown in Fig. 9b. While the x-error bars indicate the standard deviation of noise levels in pixels, the y-error bars indicate the standard deviation in AUC values of all ten classes. As a general rule, the classification improves with lower noise levels. We first note that the classification does not become perfect for any noise level and there is a significantly diminishing return in increasing the SNR beyond a level. At the other end, the ability to distinguish classes is entirely lost at levels of ca. 0.1. Performance across multiple data sets observed using our prediction model indicates that the increases demonstrated at noise levels lower than ca. 0.003 a.u. are within the variance. Hence, there is little benefit to decreasing the noise levels below ca. 0.003 a.u. for this data set, or to increasing the SNR beyond ca. 150. It must be emphasized that the model, prediction algorithm, and discriminant function are intimately linked in a non-linear manner. While this makes it impossible to predict the behavior generally of all classification approaches, this simple exercise may be conducted to determine the optimal data acquisition parameters. For our selected metrics and model, it appears that the data acquisition time can be decreased by a factor of ca. 3 without significant degradation in accuracy.

Spectral resolution

We next examined the effect of spectral resolution on the results that would be obtained using the developed algorithm. As in the previous section, the data were not re-acquired but were downsampled from acquired data using a neighbor binning procedure. Spectra from the same epithelial class pixel, at different resolutions (Fig. 10a), demonstrate the effect of downsampling on feature definition. Figure 10b demonstrates, first, that the peak-to-peak noise levels over the region remain the same with spectral resolution. As previously observed, noise is an important control in comparing spectra; the peak-to-peak noise over the same number of data points was preserved by neighbor binning. In practice, the constant-throughput spectrometer would provide a SNR (or noise level, in this case) that decreases linearly with resolution. Since most array

[2] It is noteworthy that we are examining trends in the absorbance spectra. Strictly, SNR should be measured in single beam spectra to relate rigorously to theory. It can be shown, however, that the trend will hold approximately for the absorbance spectra as well. Many practitioners advocate the use of rms SNR. We are employing peak-to-peak fluctuations over the same spectral range. Hence, the noise values we obtain will be higher but will follow the same trend.

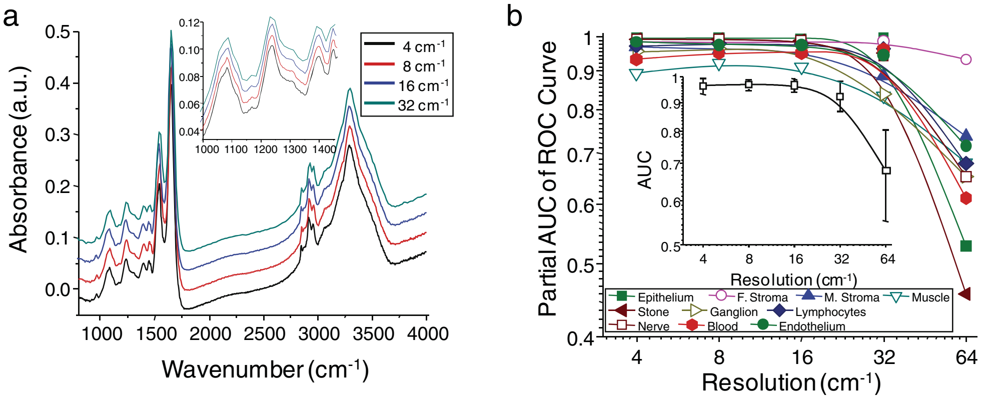


Fig. 10 **a** Spectra obtained by downsampling acquired data to different resolutions using a neighbor binning procedure. The *inset* demonstrates the effect of resolution on narrower features in the spectrum. **b** AUC values for each class and average AUC values as a function of spectral resolution demonstrate a decrease only for coarse spectral resolution

detectors can be operated with higher integration times, it is fair to assume that the time advantage in decreasing resolution would be linear. Second, the performance of the classifier is very nearly the same for finer spectral resolutions and degrades only significantly for 32 cm^{-1}. While the results may appear to be surprising, a closer analysis of the basis of the algorithms provides insight into the trends.

The classifier is based on absorbance and center of gravity measures of the peaks. It is well established that absorbance is measured accurately, provided that the FWHH of the peak is not significantly smaller than the resolution. The Ramsay resolution parameter, σ, is a useful measure that was originally developed for monochromators but has been shown to be applicable to FTIR spectrometers as well [74]. While most bands are broad and peak absorbance lower than ca. 0.7, absorbance values are not expected to be adversely impacted from the measurement process. With decreasing resolution, however, broadening within complex peaks shapes may lead to observed changes in the apparent absorption at a specific wavenumber. The change itself may not have a significant influence on the classifier performance as it depends on several such metrics. A second type of metric calculates the area under the curve. This is not expected to be impacted significantly for most peaks. The third type of metric we have used is the center of gravity of a spectral region. While spectral analyses ordinarily attempt to locate the peak position and use it as a metric, we chose the center of gravity for its sensitivity to both position and asymmetrical shape changes in complex spectral envelopes observed in biological samples. Since the classifier is based on center of gravity of a feature and not on the wavenumber of the peak maximum, it is a very robust measure that is relatively unaffected by spectral resolution or noise.

Generalization of developed algorithms to instruments and practical approaches

The characterization of classification with regard to spectrometer performance (SNR) and spectral resolution provides information to optimize parameters on one spectrometer. It is unclear, however, if the calibration would transfer to another spectrometer. We contend that the potential for a successful transfer is high as the classification process is relatively insensitive to resolution, implying that it would only be weakly sensitive to apodization or to small inaccuracies in wavelength scale. Similarly, if the SNR of acquired data is used as control, perturbations due to fixed pattern noise in focal plane array detectors or the different use of electronic filters by different manufacturers is likely to be insignificant in classifying tissue correctly. Various instrument manufacturers also set the nominal optical resolution differently in their instruments. The issue of spatial resolution, of course, is more complex. Nevertheless, any resolution setting around the wavelength-limited case will likely provide consistent results. To our knowledge, there has been no comparison yet of classifier performance across mid-IR FTIR imaging spectrometers using algorithms developed on one specific instrument. The developed protocol provides for such a framework and detailed results are awaited from on-going work [75].

Outlook and prospects

An exciting period in imaging tissues spectroscopically with low power, optical microscopy-comparable resolution is emerging. Considerable work, however, needs to be accomplished before this idea can become a clinical reality. An ultimate goal of such studies is to provide a key technology for emerging molecular pathology. The approach promises greatly reduced error rates, automation, and economic benefits in current pathology practice. Looking to the future, chemical imaging approaches will be employed for diagnosing cancers in pre-malignant stages prior to their apparent changes observable by conventional means, predicting the prognosis of the lesion and intra-operative imaging in real-time. Fundamental studies in drug discovery and mechanisms of molecular interactions are further examples that would be enabled by progress in this area. Doubtless, exciting applications lie ahead and progress is rapidly being made towards practical applications but much work needs to be done to carefully apply this powerful technology to multiple aspects of pathology. Success in this endeavor promises to change the practice of pathology radically and alter the clinical management of cancer patients.

Acknowledgement The author would like to acknowledge collaborators over the years, especially Dr. Stephen M. Hewitt and Dr. Ira W. Levin of the National Institutes of Health, for numerous useful discussions and guidance. Discussions and help from Dr. Daniel Fernandez during the formative years of this work are also appreciated. Funding for this work was provided in part by University of Illinois Research Board and by the Department of Defense Prostate Cancer Research Program. This work was also funded in part by the National Center for Supercomputing Applications and the University of Illinois, under the auspices of the NCSA/UIUC faculty fellows program.

References

1. Woolf SH (1995) N Engl J Med 333:1401–1405
2. Humphrey PA (2003) Prostate pathology. American Society for Clinical Pathology, Chicago
3. Partin AW, Mangold LA, Lamm DM, Walsh PC, Epstein JI, Pearson JD (2001) Urology 58:843–848
4. De La Taille A, Viellefond A, Berger N, Boucher E, De Fromont M, Fondimare A, Molinié V, Piron D, Sibony M, Staroz F, Triller M, Peltier E, Thiounn N, Rubin MA (2003) Hum Pathol 34:444–449
5. Levin IW, Bhargava R (2005) Annu Rev Phys Chem 56:429–474
6. Navratil M, Mabbott GA, Arriaga EA (2006) Anal Chem 78:4005–4019
7. Caprioli RM, Farmer TB, Gile J (1997) Anal Chem 69:4751–4760
8. Chaurand P, Schwartz SA, Billheimer D, Xu BGJ, Crecelius A, Caprioli RM (2004) Anal Chem 76:1145–1155
9. Kurhanewicz J, Vigneron DB, Hricak H, Narayan P, Carroll P, Nelson S (1996) Radiology 198:795–805
10. Lewis EN, Gorbach AM, Marcott C, Levin IW (1996) Appl Spec 50:263–269
11. Diem M, Romeo M, Boydston-White S, Miljkovic M, Matthaus C (2004) Analyst 129:880–885
12. Mendelsohn R, Paschalis EP, Boskey AL (1999) J Biomed Opt 4:14–21
13. Kidder LH, Kalasinsky VF, Luke JL, Levin IW, Lewis EN (1997) Nat Medicine 3:235–237
14. Ellis DI, Goodacre R (2006) Analyst 131:875–885
15. Bhargava R, Levin IW (eds) (2005) Spectrochemical analysis using infrared multichannel detectors. Blackwell, Oxford
16. Petrich W (2001) Appl Spectrosc Rev 36(2):181–237
17. Andrus PG (2006) Tech Cancer Res Treat 5:157–167
18. Krafft C, Sergo V (2006) Spectroscopy 20:195–218
19. Petibois C, Deleris G (2006) Trends Biotechnol 24:455–462
20. Walsh MJ, German MJ, Singh M, Pollock HM, Hammiche A, Kyrgiou M, Stringfellow HF, Paraskevaidis E, Martin-Hirsh PL, Martin FL (2007) Cancer Lett 246:1–11
21. Keith FN, Bhargava R (2007) Tech Cancer Res Treat (submitted)
22. Gazi E, Dwyer J, Gardner P, Ghanbari-Siakhali A, Wade AP, Myan J, Lockyer NP, Vickerman JC, Clarke NW, Shanks JH, Hart C, Brown M (2003) J Pathology 201:99–108
23. Gazi E, Baker M, Dwyer J, Lockyer NP, Gardner P, Shanks JH, Reeve RS, Hart C, Clarke NW, Brown M (2006) Eur Urol 50:750–761
24. Harvey TJ, Henderson A, Gazi E, Clarke NW, Brown M, Faria EC, Snook RD, Gardner P (2007) Analyst 132:292–295
25. Paluszkiewicz C, Kwiatek WM, Banas A, Kisiel A, Marcelli A, Piccinini A (2007) Vib Spectrosc 43:237–242
26. Fernandez DC, Bhargava R, Hewitt SM, Levin IW (2005) Nat Biotechnol 23:469–474
27. German MJ, Hammiche A, Ragavan N, Tobin MJ, Cooper LJ, Matanhelia SS, Hindley AC, Nicholson CM, Fullwood NJ, Pollock HM, Martin FL (2006) Biophys J 90:3783–3795
28. Gazi E, Dwyer J, Lockyer NP, Miyan J, Gardner P, Hart CA, Brown MD, Clarke NW (2005) Vib Spectrosc 38:193–201
29. Bhargava R, Hewitt SM, Levin IW (2007) Nat Biotechnol 25:31–33
30. Srinivasan G, Bhargava R (2007) Spectroscopy 22:30–43
31. Bhargava R, Fernandez DC, Hewitt SM, Levin IW (2006) Biochim Biophys Acta Biomembr 1758:830–845
32. Swets JA (1988) Science 240:1285–1293
33. Lasch P, Naumann D (2006) Biochim Biophys Acta 1758:814–829
34. Jackson M, Choo LP, Watson PH, Halliday WC, Mantsch HH (1995) Biochim Biophys Acta 1270:1–6
35. Sommer AJ, Katon JE (1991) Appl Spectrosc 45:1633–1640
36. Carr GL (2001) Rev Sci Inst 72:1613–1619
37. Bhargava R, Wang SQ, Koenig JL (1998) Appl Spectrosc 52:323–328
38. Budevska BO (2000) Vib Spectrosc 24:37–45
39. Romeo M, Diem M (2005) Vib Spectrosc 38:129–132
40. Jackson M (2004) Faraday Discuss 126:1–18
41. Norris KP (1954) J Sci Inst 31:284–287
42. Rousch PB (ed) (1985) The design, sample handling, and applications of infrared microscopes. ASTM STP 949, American Society for Testing and Materials, Philadelphia
43. Kwiatkoski JM, Reffner JA (1987) Nature 328:837–838
44. Koenig JL (1999) Spectroscopy of polymers, 2nd edn. Elsevier, New York
45. Bartick EG, Tungol MW, Reffner JA (1994) Anal Chim Acta 288:35–42
46. Wetzel DA, LeVine SM (1999) Science 285:1224–1225
47. Gremlich H-U, Yan B (eds) (2000) Infrared and Raman spectroscopy of biological materials (practical spectroscopy). Marcel Dekker, New York
48. Bhargava R, Wall BG, Koenig JL (2000) Appl Spectrosc 54:470–474
49. Vobornik D, Margaritondo G, Sanghera JS, Thielen P, Aggarwal ID, Ivanov B, Miller JK, Haglund R, Tolk NH, Congiu-Castellano A, Rizzo MA, Piston DW, Somma F, Baldacchini G, Bonfigli F, Marolo T, Flora F, Montereali RM, Faenov A, Pikuz T, Longo G, Mussi V, Generosi R, Luce M, Perfetti P, Cricenti A (2004) Infrared Phys Tech 45:409–416
50. Hirschfeld T (1979) Appl Spectrosc 33:525–527
51. Wetzel DL (2002) Vib Spectrosc 29:183–189
52. Carter MR, Bennett CL, Fields DJ, Hernandez J (1995) Proc SPIE 2480:380–386
53. Lewis EN, Treado PJ, Reeder RC, Story GM, Dowrey AE, Marcott C, Levin IW (1995) Anal Chem 67:3377–3381
54. Colarusso P, Kidder LH, Levin IW, Fraser JC, Arens JF, Lewis EN (1998) Appl Spectrosc 52:106A–120A
55. Snively CM, Koenig JL (1999) Appl Spectrosc 53:170–177
56. Bhargava R, Levin IW (2001) Anal Chem 73:5157–5167
57. Ransohoff DF (2004) Nat Rev Cancer 4:309–314
58. Bhargava R, Levin IW (eds) (2005) Spectrochemical analysis using infrared multichannel detectors. Blackwell , Oxford, pp 56–84
59. Various contributors (2006) Biochim Biophys Acta Biomembr 1758

60. Wood BR, Chiriboga L, Yee H, Quinn MA, McNaughton D, Diem M (2004) Gynecol Oncol 93:59–68
61. Malins DC, Polissar NL, Nishikida K, Holmes EH, Gardner HS, Gunselman SJ (1995) Cancer 75:503–517
62. Boydston-White S, Gopen T, Houser S, Bargonetti J, Diem M (1999) Biospectroscopy 5:219–227
63. Shaw RA, Guijon FB, Paraskevas V, Ying SL, Mantsch HH (1999) Anal Quant Cytol 21:292–302
64. Mansfield JR, McIntosh LM, Crowson AN, Mantsch, HH, Jackson, M (1999) Appl Spectrosc 53:1323–1333
65. McIntosh LM, Jackson M, Mantsch HH, Stranc MF, Pilavdzic D, Crowson AN (1999) J Invest Dermatol 112:951–956
66. Kononen J, Bubendorf L, Kallioniemi A, Barlund M, Schraml P, Leighton S, Torhorst J, Mihatsch MJ, Sauter G, Kallioniemi OP (1998) Nat Med 4:844–847
67. Camp RL, Charette LA, Rimm DL (2000) Lab Invest 80:1943–1949
68. Paluszkiewicz C, Kwiatek WM, Banas A, Kisiel A, Marcelli A, Piccinini M (2007) Vib Spectrosc 43(1):237–242
69. Benjamini Y, Hochberg Y (1995) J R Stat Soc Ser B 57:289–300
70. Pawitan Y, Michiels S, Koschielny S, Gusnanto A, Ploner A (2005) Bioinformatics 21:3017–3024
71. Stone N, Kendall C, Smith J, Crow P, Barr H (2004) Faraday Diss 126:141–157
72. Bhargava R, Wang SQ, Koenig JL (2000) Appl Spectrosc 54:486–495
73. Bhargava R, Wang SQ, Koenig JL (2000) Appl Spectrosc 54:1690–1706
74. Anderson RJ, Griffiths PR (1975) Anal Chem 47:2339–2347
75. Llora X, Reddy RK, Bhargava R (in preparation)

Nat Comput
DOI 10.1007/s11047-007-9056-6

Observer-invariant histopathology using genetics-based machine learning

Xavier Llorà · Anusha Priya · Rohit Bhargava

Abstract Prostate cancer accounts for one-third of noncutaneous cancers diagnosed in US men and is a leading cause of cancer-related death. Advances in Fourier transform infrared spectroscopic imaging now provide very large data sets describing both the structural and local chemical properties of cells within prostate tissue. Uniting spectroscopic imaging data and computer-aided diagnoses (CADx), our long term goal is to provide a new approach to pathology by automating the recognition of cancer in complex tissue. The first step toward the creation of such CADx tools requires mechanisms for automatically learning to classify tissue types—a key step on the diagnosis process. Here we demonstrate that genetics-based machine learning (GBML) can be used to approach such a problem. However, to efficiently analyze this problem there is a need to develop efficient and scalable GBML implementations that are able to process very large data sets. In this paper, we propose and validate an efficient GBML technique—`NAX`—based on an incremental genetics-based rule learner. `NAX` exploits massive parallelisms via the message passing interface (MPI) and efficient rule-matching using hardware-implemented operations. Results demonstrate that `NAX` is capable of performing prostate tissue classification efficiently, making a compelling case for using GBML implementations as efficient and powerful tools for biomedical image processing.

X. Llorà (✉)
National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign,
1205 W. Clark Street, Urbana, IL 61801, USA
e-mail: xllora@uiuc.edu

A. Priya · R. Bhargava
Department of Bioengineering, University of Illinois at Urbana-Champaign, 1304 W. Springfield Ave.,
Urbana, IL 61801, USA

A. Priya
e-mail: priya@uiuc.edu

R. Bhargava
e-mail: rxb@uiuc.edu

R. Bhargava
Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign,
405 N. Mathews Ave., Urbana, IL 61801, USA

Keywords Observer-invariant histopathology · Genetics-based machine learning · Learning Classifier Systems · Hardware acceleration · Vector instruction · SSE2 · MPI · Massive parallelism

1 Introduction

Pathologist opinion of structures in stained tissue is the definitive diagnosis for almost all cancers and provides critical input for therapy. In particular, prostate cancer accounts for one-third of noncutaneous cancers diagnosed in US men. Hence, it is, appropriately, the subject of heightened public awareness and widespread screening. If prostate-specific antigen (PSA) or digital rectal screens are abnormal, a biopsy is needed to definitively detect or rule out cancer. Pathologic status of biopsied tissue not only forms the definitive diagnosis but constitutes an important cornerstone of therapy and prognosis. There is, however, a need to add useful information to diagnoses and to introduce new technologies that allow economical cancer detection to focus limited healthcare resources. In pathology practice, widespread screening results in a large workload of biopsied men, in turn, placing a increasing demand on services. Operator fatigue is well documented and guidelines limit the workload and rate of examination of samples by a single operator. Importantly, newly detected cancers are increasingly moderate grade tumors in which pathologist opinion variation complicates decision-making.

For the reasons above, there is an urgent need for automated and objective pathology tools. We have sought to address these requirements through novel Fourier transform infrared (FTIR) spectroscopy-based, computer-aided diagnoses for prostate cancer and develop the required microscopy and software tools to enable its application. FTIR spectroscopic imaging is a new technique that combines the spatial specificity of optical microscopy and the biochemical content of spectroscopy. As opposed to thermal infrared imaging, FTIR imaging measures the absorption properties of tissue through a spectrum consisting of (typically) 1024–2048 wavelength elements per pixel. Since IR spectra reflect the molecular composition of the tissue, image contrast arises from differences in endogenous chemical species. As opposed to visible microscopy of stained tissue that requires a human eye to detect changes, numerical computation is required to extract information from IR spectra of unstained tissue. Extracted information, based on a computer algorithm, is inherently objective and automated (Lattouf and Saad 2002; Fernandez et al. 2005; Levin and Bhargava 2005; Bhargava et al. 2006).

Uniting spectroscopic imaging data and computer-aided diagnoses (CADx), we seek to provide a new approach to pathology by automating the recognition of cancer in complex tissue. This is an exciting paradigm in which disease diagnoses are objective and reproducible; yet do not require any specialized reagents or human intervention. The first step toward the creation of such CADx tools requires mechanisms for reliable and automated tissue type classification. In this paper we demonstrate how genetics-based machine learning tools can achieve such a goal. Interpretability of the learned models and efficient processing of very large data sets have lead us to rule-based models—easy to interpret—and genetics-based machine learning—inherent massively parallel methods with the required scalability properties to address very large data sets. We present the method and the efficiency enhancement techniques proposed to address automated tissues classification. When pushed beyond the relatively small problems traditionally used to test such methods, an need for efficient and scalable implementations becomes a key research topic that needs to be addressed. We designed the proposed a technique with such constraints in mind. A modified version of an incremental genetics-based rule learner that exploits massive parallelisms—via the message passing interface (MPI)—and efficient rule-matching using hardware-oriented operations. We name this system `NAX`. `NAX` is compared to traditional and genetics-based machine learning techniques on an array of publicly available data sets. We also report the initial results achieved using the proposed technique when classifying prostate tissue.

The remainder of the paper is structured as follows. We present an overview of the problem addressed in Sect. 2, paying special attention to tissue classification. We discuss in Sect. 3 the hurdles that traditional genetics-based machine learning implementations face when applied to very large data sets. Section 4 presents our solution to those hurdles. We also describe the incremental rule learner proposed for tissue classification. Last, we summarize results on publicly-available data sets and the preliminary results for tissue classification on a prostate tissue microarray in Sect. 5. Finally, in Sect. 6, we present conclusions and further work.

2 Biomedical imaging and data mining

This section presents an overview of the problem addressed in this paper. We first introduce infrared spectroscopic imaging as a potentially powerful tool for cancer diagnosis and prognosis. Then, we explore the protocols that provide raw high-quality data that for data mining. Finally, we conclude by focusing on the key task, tissue classification, by focusing on prostate tissue.

2.1 Infrared spectroscopy and imaging for cancer diagnosis and prognosis

Infrared spectroscopy is a well-established molecular technique and is widely used in chemical analyses. The fundamental principle governing the response of any material is that the vibrational modes of molecules are resonant in energy with photons in the mid-infrared region (2–14 mm) of the electromagnetic spectrum. Hence, when photons of energy that are resonant with the material's molecular composition are incident, a number are absorbed. The number absorbed is directly proportion to the number of chemical species that are excited. Hence, any material has a characteristic frequency-dependent absorption profile called a spectrum. An infrared spectrum is often termed the "optical fingerprint" of a material as it can help uniquely identify molecular composition—see Fig. 1.

Researchers, including us, have contributed to develop an imaging version of spectroscopy that is essentially similar to an optical microscope. In this mode of spectroscopy, images are acquired in the manner of optical microscopy with one important difference. Instead of measuring the intensity of three colors for a visible image, several thousand intensity values are acquired at each pixel in the image as a function of wavelength (spectrum at each pixel). The resulting data set is three dimensional (2 spatial and 1 spectral indices) consisting typically of a size $256 \times 256 \times 1024$, but extending to sizes such as $3500 \times 3500 \times 2048$. Since each data point is stored as a 16-bit number, the data size typically runs into several tens to hundreds of gigabytes.

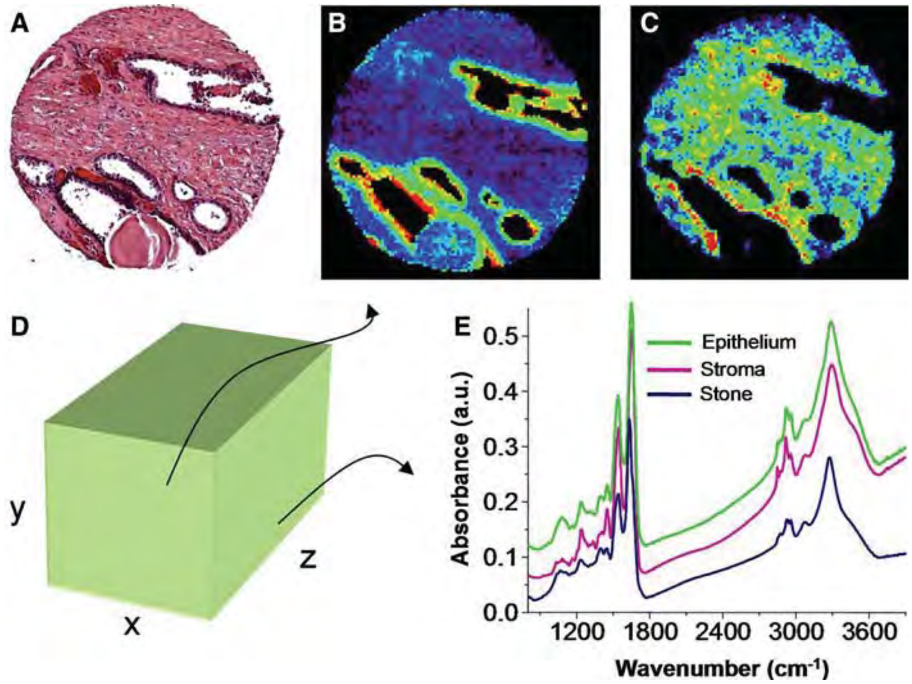


Fig. 1 Conventional staining and automated recognition by chemical imaging. (**A**) Typical H&E stained sample, in which structures are deduced from experience by a human. Highlights of specific regions in the manner of H&E is possible using FTIR imaging without stains. (**B**) Absorption at 1080 cm^{-1} commonly attributed to nucleic acids and (**C**) to proteins of the stroma. The data obtained is 3 dimensional (**D**) from which spectra (**E**) or images at specific spectral features may be plotted

2.2 Mining the spectra: Two sequential problems

Though the continued development of fast FTIR microspectroscopy represents an exciting opportunity for pathology, handling the resultant data and rapidly providing classifications remains a critical challenge. First, the sheer volume of data—potentially larger than 10 GB a day—represents an organizational and retrieval challenge. Next, extraction of useful information in short time periods requires the formulation of optimal protocols. Third, the automated cancer segmentation problem is very complex and offers a number of routes and levels of data that need to be analyzed to determine the optimal approach for application in a laboratory.

The typical application is the need to extract information from the data set such that it is clinically relevant. Hence, the output of the data mining algorithm to be developed is well-bounded and clearly defined. For example, in the prostate there are two levels of interest. In the first level, the pathologist examines the tissue to determine if there are any epithelial cells. Since more than 95% of prostate cancers arise in epithelial cells, transformations in this class of cells forms the diagnostic basis and a primary determinant of therapy. Other cell types of interest are lymphocytes that may indicate inflammation, blood vessel density that may indicate the development of new blood supply indicative of cancer growth and nerves that may be invaded by cancer cells. Hence, any automated approach to pathology must first identify cell types accurately. The second step in pathology follows. Once epithelial cells are located, their spatial patterns are indicative of disease states. In our imaging approach, we can identify both spatial patterns as well as chemical patterns in epithelial cells. Hence, the task would be to use either or both to classify disease. In this paper, we focus only on the accurate identification/classification of tissue types as the first step of the path that leads to obtaining the correct pixels of epithelium.

2.3 Tissue classification for prostate arrays

Prostate tissue is structurally complex, consisting primarily of glandular ducts lined by epithelial cells and supported by heterogeneous stroma. This tissue also contains blood vessels, blood, nerves, ganglion cells, lymphocytes and stones (which are comprised of luminal secretions of cellular debris) that organize into structure measuring from tens to hundreds of microns. These structures are readily observable within stained tissue using bright-field microscopy at low to medium magnifications. Hence, in applying FTIR imaging (Levin and Bhargava 2005), we obtain the common structural detail employed clinically and, additionally, spectral information indicative of tissue biochemistry. As histologic classes contain identical chemical components, infrared vibrational spectra are similar but reveal small differences in specific absorbance features. The technique proposed by Fernandez et al. (2005) examines each cell types' spectra and transforms each spectrum into a vector of describing features—usually around the hundreds. A complete description of this process is beyond the scope of this paper and can be found elsewhere (Fernandez et al. 2005). Each pixel (cell present in the slice of micro array under analysis) has an assigned spatial position in the array while the tissue type is assigned by a highly experienced pathologist. Thus, the tissue classification can be cast into a supervised classification problem (Mitchell 1997), where all the attributes are real-valued and the class is the tissue type—ten classes: *ephithelium*, *fibrous stroma*, *mixed stroma*, *muscle*, *stone*, *lymphocytes*, *endothelium*, *nerve*, *ganglion*, and *blood*. Figure 2 presents tissue types that can be assigned by examining a stained image obtained, after the FTIR microsprectroscopy on unstained tissue,by the pathologist. Each marked pixel in the image becomes a training example; hence, the usual smallest data set is around hundreds of thousand records per array.

3 Larger, bigger, and faster genetics-based machine learning

Bernadó et al. (2001) presented a first empirical comparison between genetics-based machine learning techniques (GBML) and traditional machine learning approached. The authors reported that GBML techniques were as competent as traditional techniques. Later, Bacardit and Butz (2006) repeated the analysis, obtaining similar results. Most of the experiments presented on both papers used publicly available data sets provided by the *University of California at Irvine* repository (Merz and Murphy 1998). Most of the data sets are defined over tens of features and up to few thousands of records—in the larger cases. However, a key property of GBML approaches is its intrinsic massive parallelism and scalability properties. Cantú-Paz (2000) presented how efficient and accurate genetics algorithms could be assembled, and Llorà (2002) presented how such algorithms can be efficiently used for machine learning and data mining. However, there are elements that need to be revisited when we want to efficiently apply GBML techniques to large data sets such as the one described in the previous section.

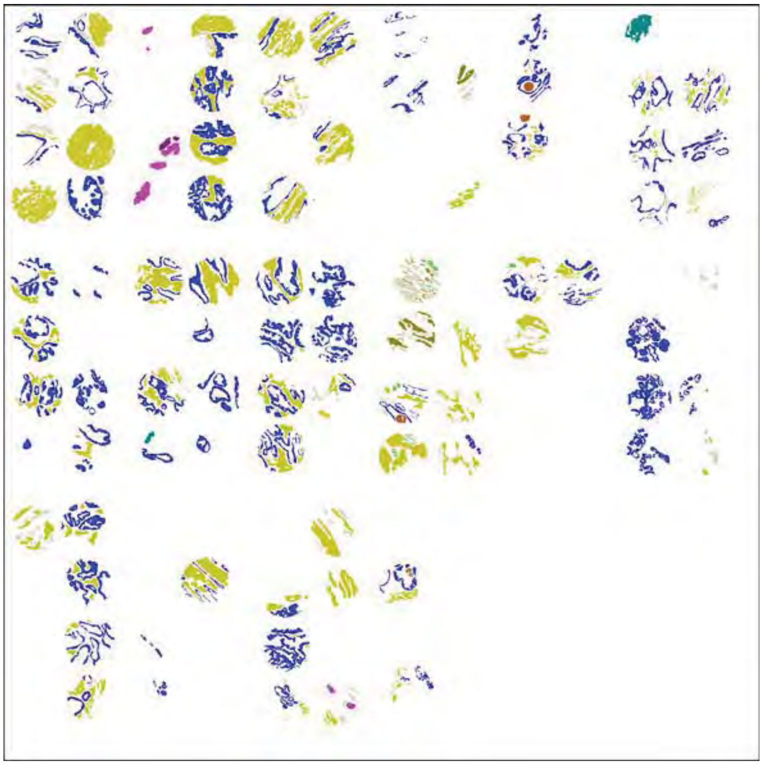


Fig. 2 The figure presents the tissue labeling provided by a pathologist biopsy section of human prostate tissue. Each spot represents the section of a needle. Different colors represent different tissue types

The GBML techniques require evaluating candidate solutions against the original data set matching the candidate solutions (e.g., rules, decision trees, prototypes) against all the instances in the data set. Regardless of the flavor used, Llorà and Sastry (2006) showed that, as the problem grows, rule matching governs the execution time. For small data sets (teens of attributes and few thousands of records) the matching process takes more than 85% of the overall execution time marginalizing the contribution of the other genetic operators. This number increases to 98% and above, when we move to data sets with few hundreds of attributes and few hundred thousands of records. More than 98% of the time is spent evaluating candidate solutions. Each evaluation can be computed in parallel. Moreover, the evaluation process may also be parallelized on very large data sets by splitting and distributing the data across the computational resources. A detailed description of the parallelization alternatives of GBML techniques can be found elsewhere (Llorà 2002).

Currently available off-the-shelf GBML methods and software distributions (Barry and Drugow-itsch 1997; Llorà 2006) do not usually target large data sets. The two main bottlenecks are large memory footprints and sequential-processing oriented processes. Generally speaking, they were designed to run on single processor machines with enough memory to fit the entire data set. Hence, designers did not paying much attention to the memory footprint required to store the data set—usually completely loaded into memory and the population of candidate solutions. These large complex structures were geared to facilitate the programming effort, but they are not designed toward the efficient evaluation of the candidate solutions. However, efforts have been made to push GBML methods into domains which require processing large data sets. Three different works need to be mentioned here. Flockhart (1995) proposed and implemented GA-MINER, one of the earliest effort to create data mining systems based on GBML systems that scale across symmetric multi-processors and massively parallel multi-processors. Flockhart (1995) reviewed different encoding and parallelization schemes and conducted proper scalability studies. Llorà (2002) explored how fine-grained parallel genetic algorithms could become efficient models for data mining. Theoretical analysis of performance and scalability were developed and validated with proper simulations. Recently, Llorà and Sastry (2006) explored how current hardware can efficiently speed up rule matching against large data sets. These three approaches are the basis of the incremental rule learning proposed in the next section to approach very large data sets.

Another important issue in real-world problems is the class distribution. Usually most real problems have a clear class imbalance. Recently, Orriols-Puig and Bernadó-Mansilla (2006) have revisited this issue, showing how GBML techniques successfully learn and maintain proper descriptions for those minority classes. If not designed properly, descriptions of majority classes will tend to govern the learned models, starving the description of minority classes. Prostate tissue classification is a clear example of extreme class imbalance. Figure 3 presents the tissue type class distribution. The smaller tissue type has 64 records, where as the larger classes have several tens of thousands records. hence, the developed approaches must account for class size variation.

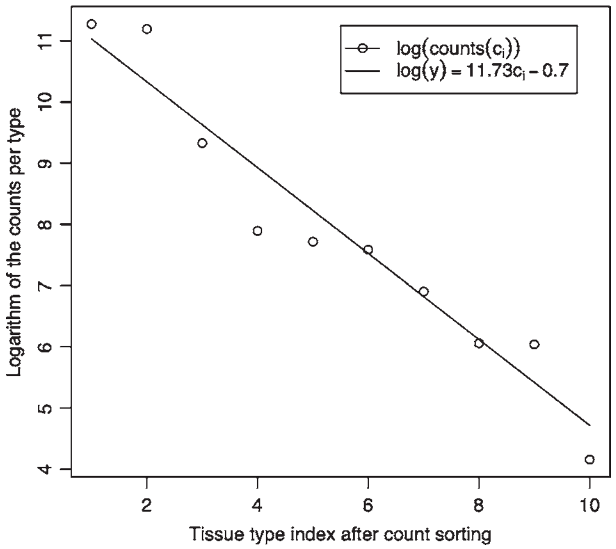


Fig. 3 Figure shows the tissue class distribution. Once the classes are reordered according to their frequency in the data set, we can easily appreciate the extreme imbalance—the smaller tissue type has 64 records, where as the larger classes have several tens of thousands records

4 The road to tractability

We describe in this section the steps we took to design a GBML method (NAX) able to deal with very large data sets with class imbalance. NAX evolves, one at a time, maximally general and maximally accurate rules. Then, the covered instance are removed and another maximally general and maximally general rule is evolved and added to the previously stored one forming a decision list. This process continues until no uncovered instances are left—this process is also referred as the sequential covering procedure (Cordón et al. 2001). Llorà et al. (2005) showed that maximally general and maximally accurate rules (Wilson 1995) could also be evolved using Pittsburgh-style Learning Classifier Systems. Later, Llorà et al. (2007) showed that competent genetic algorithms (Goldberg 2002) evolve such rules quickly, reliably, and accurately. The rest of this section describes (1) efficient implementation techniques to deal with very large data sets, (2) the impact of class imbalance, and (3) the NAX algorithm proposed.

4.1 Efficient implementations

As introduced earlier, when dealing with very large data sets, and regardless of the flavor of the GBML technique used, we may spend up to 98% of the computational cycles trying to match rules to the original data set (Llorà and Sastry 2006). Each solution evaluation is independent of each other and, hence, it can be computed in parallel. Moreover, even the matching nature of a rule—the representation we will use from now on—is highly parallel, since conditions require performing simultaneous checks against different attributes per record. Thus, efficient implementation can take advantage of parallelizing both elements.

4.1.1 Exploiting the hardware acceleration

Recently, multimedia and scientific applications have pushed CPU manufactures to include support for vector instructions again in their processors. Both applications areas require heavy calculations based on vector arithmetic. Simple vector operations such as *add* or *product* are repeated over and over. During 1980s and 1990s supercomputers, such as Cray machines, were able to issue hardware instructions that enabled basic vector arithmetics. A more constrained scheme, however, has made its way into general-purpose processors thanks to the push of multimedia and scientific applications. Main chip manufactures—IBM, Intel, and AMD—have introduced vector instruction sets—Altivec, SSE3, and 3DNow$^+$—that allow vector operations over packs of 128 bits by hardware. We will focus on a subset of instructions that are able to deal with floating point vectors. This subset of instructions manipulate groups of four floating-point numbers. These instructions are the basis of the fast rule matching mechanism proposed.

Our goal is to evolve a set of rules that correctly classifies the current data set rom prostate tissue. Using a knowledge representation based on rules allows us to inspect the learned model, gaining insight into the biological problem as well. All the attributes of the domain are real-value and the conditions of the rules need to be able to express conditions in a $\Re^n$ spaces. We use a similar rule encoding to the one proposed by Wilson (2000b)—a variation of the original work proposed by Wilson (2000a) and later reviewed by Stone and Bull (2003)—and widely used in the GBML community. Rules express the conjunction of tests across attributes. Each test may be defined in multiple flavors but, without loss of generality, we picked a simple interval based one. A simple example of an *if-then* rule, could be expressed as follows:

$$1.0 \leq a_0 \leq 2.3 \wedge \cdots \wedge 10.0 \leq a_n \leq 23 \rightarrow c_1 \tag{1}$$

Where the condition is the conjunction of the different attribute tests and the outcome is the predicted class—a tissue type. We also allow a special condition—`don't care` —which just always returns `true` , allowing condition generalization. The rule below illustrates an example of a generalized rule.

$$1.0 \leq a_0 \leq 2.3 \wedge -3.0 \leq a_3 \leq 2 \longrightarrow c_1 \tag{2}$$

All attributes except a_0 and a_3 were marked as `don't care`.

Each condition can be encoded using 2 floating-point numbers per condition, where α_i contains the lower bound of the condition and ω_i its upper bound. Thus, the condition $\alpha_i \leq a_0 \leq \omega_i$ just requires to store the two floating-point numbers. For efficiency reasons we store them in two separate vectors, on containing the lower bounds and the other containing the upper bounds. The position in a vector indicates the attribute being tested. The `don't care` condition is simply encoded as $\alpha_i > \omega_i$ and, hence, we do not need to store any extra information.

Matching a rule requires performing the individual condition tests before the final *and* operation can be computed. Vector instruction sets improve the performance of this process by performing four operations at once. Actually, this process may be regarded as four parallel running pipelines. The process can be further improved by stopping the matching process when one test fails—since that will turn the condition into false.

Figure 4 presents a `C` implementation the proposed hardware-supported rule matching. The code assumes that the two vectors containing the upper and lower bounds are provided and records are stored in a two dimensional matrix. Figure 5 presents the vectorized implementation of the code presented in Fig. 4 using SSE2 instructions. Exploiting the hardware available can speed between 3 and 3.5 times the matching process, as also shown elsewhere (Llorà and Sastry 2006).

4.1.2 Massive parallelism

Since most of the time is spent on the evaluation of candidate rules when dealing with large data sets, our next goal was to find a parallelization model that could take advantage of this peculiarity. Due the quasi embarrassing parallel (Grama et al. 2003) nature of the candidate rule evaluation, we designed a coarse-grain parallel model for distributing the evaluation load. Cantú-Paz (2000) proposed several schemes, showing the importance of the trade-off between computation time and time spent communicating. When designing the parallel model, we focused on minimizing the communication cost. Usually, a feasible solution could be a master/slave one—the computation time is much larger than the communication time. However, GBML approaches tend to use rather large populations, forcing us to send rule sets to the evaluation slaves and collect the resulting fitness. These schemes also increment the sequential sections that cannot be parallelized, threatening the overall speedup of the parallel implementation as a result of Ambdhals law (Amdahl 1967).

To minimize such communication cost, each processor runs an identical `NAX` algorithm. They are all seeded in the same manner, hence, performing the same genetic operations and only differing in the portion of the population being evaluated. Thus, the population is

```
 1. void match_seq_rule_set ( RuleSet * rs, InstanceSet is, int iDim, int iRows ) {
 2.    int i,j,k,iCnt,iClsIdx,iGround,iPred;
 3.    register int iMatcheable;
 4.    Instance ins;
 5.
 6.    iClsIdx = rs->iCorrectedDim;
 7.    clean_fitness_rules_set(rs);
 8.    for ( i=0 ; i<iRows ; i++ ) {
 9.       ins = is[i];
10.       iPred=-1;
11.       for ( j=0 ; iPred==-1 && j<rs->iLen ; j++ ) {
12.          iMatcheable = 1;
13.          for ( iCnt=0,k=j*(rs->iCorrectedDim+VBSIF) ;
14.                iMatcheable && k<j*(rs->iCorrectedDim+VBSIF)+rs->iDim ;
15.                k++,iCnt++ ) {
16.             iMatcheable = iMatcheable &&
17.                           !( (rs->pfLB[k]<=rs->pfUB[k]) &&
18.                           ( ins[iCnt]<rs->pfLB[k] || ins[iCnt]>rs->pfUB[k]));
19.          }
20.          if ( iMatcheable )
21.             iPred = rs->pfLB[j*(rs->iCorrectedDim+VBSIF)+rs->iCorrectedDim];
22.       }
23.       iPred = (iPred==-1)?rs->iClasses:iPred;
24.       iGround=(int)ins[iClsIdx];
25.       rs->pConfMat[iGround][iPred]++;
26.    }
27. }
```

Fig. 4 This figure presents a sequential implementation of the rule matched process in C . A rule set is match against a data set. Lines 16, 17, and 18 implement the condition test for one attribute. The implementation also computes the confusion matrix that contains the ground truth versus predicted class

treated as collection of chunks where each processor evaluates its own assigned chunk, sharing the fitness of the individuals in its chunk with the rest of the processors. Fitness can be encapsulated and broadcasted maximizing the occupation of the underlying packing frames used by the network infrastructure. Moreover, this approach also removes the need for sending the actual rules back and forth between processors—as a master/slave approach would require—thus, minimizing the communication to the bare minimum—the fitness. Figure 6 presents a conceptual scheme of the parallel architecture of NAX.

To implement the model presented in Fig. 6, we used C and a *message passing interface* (MPI)—we used the OpenMPI implementation (Gabriel et al. 2004). Figure 7 shows the code in charge of the parallel evaluation. Each processor computes which individuals are assigned to it. Then it computes the fitness and, finally, it just broadcast the computed fitness. The rest of the process is left untouched, and besides the cooperative evaluation, all the processors end generating the same evolutionary trace.

4.2 Rule sets as individuals

One main characteristic of the so-called Pittsburgh-style learning classifier systems—a particular type of GBML—is that individuals encode a rule set (Goldberg 1989; Llorà and Garrell 2001; Goldberg 2002). Thus, evolutionary mechanisms directly recombine one rule set against another one. For classification tasks of moderate complexity, the rule sets are

```
 1. #define VEC_MATCH(vecFLB,fLB,vecFUB,fUB,vecINS,fIN,vecTmp,vecOne,vecRes) {\
 2.     vecFLB = _mm_load_ps(fLB);\
 3.     vecFUB = _mm_load_ps(fUB);\
 4.     vecINS = _mm_load_ps(fIN);\
 5.     \
 6.     vecRes = (__m128i)_mm_cmpgt_ps(vecFUB,vecFLB);\
 7.     vecTmp = _mm_or_si128(\
 8.                     (__m128i)_mm_cmpgt_ps(vecFLB,vecINS),\
 9.                     (__m128i)_mm_cmpgt_ps(vecINS,vecFUB)\
10.                   );\
11.     vecRes = _mm_andnot_si128(_mm_and_si128(vecRes,vecTmp),vecOne);\
12. }
13.
14. void match_rule_set ( RuleSet * rs, InstanceSet is, int iDim, int iRows ) {
15.     int i,j,k,iCnt,iClsIdx,iGround,iPred;
16.     register int iMatcheable;
17.     Instance ins;
18.
19.     __m128i vecRes,vecTmp,vecOne;
20.     __m128 vecFLB,vecFUB,vecINS;
21.
22.     vecOne = (__m128i){-1,-1};
23.
24.     iClsIdx = rs->iCorrectedDim;
25.     clean_fitness_rules_set(rs);
26.     for ( i=0 ; i<iRows ; i++ ) {
27.         // Classify the instance
28.         ins = is[i];
29.         iPred=-1;
30.         for ( j=0 ; iPred==-1 && j<rs->iLen ; j++ ) {
31.             iMatcheable = 1;
32.             for ( iCnt=0,k=j*(rs->iCorrectedDim+VBSIF) ;
33.                   iMatcheable && k<j*(rs->iCorrectedDim+VBSIF)+rs->iDim ;
34.                   k+=VBSIF,iCnt+=VBSIF ) {
35.                 VEC_MATCH(vecFLB,&(rs->pfLB[k]),
36.                           vecFUB,&(rs->pfUB[k]),
37.                           vecINS,&(ins[iCnt]),vecTmp,vecOne,vecRes);
38.                 iMatcheable = 0xFFFF==_mm_movemask_epi8(vecRes);
39.             }
40.         if ( iMatcheable )
41.             iPred = rs->pfLB[j*(rs->iCorrectedDim+VBSIF)+rs->iCorrectedDim];
42.         iPred = (iPred==-1)?rs->iClasses:iPred;
43.         iGround=(int)ins[iClsIdx];
44.         rs->pConfMat[iGround][iPred]++;
45.     }
46. }
```

Fig. 5 This figure presents a vectorized implementation of the rule matching process presented in Fig. 4. Lines 1–12 implement the parallelized test against four attributes using vector instructions. The code is written using C intrinsics for SSE2 compatible architectures. This code runs on P4 or newer Intel processors and Opteron or Athlon 64 AMD processors

not large. However, for complex problems, the potential number of required rules to ensure proper classification may need large amounts of memory that become prohibitive. The requirements increase even further in the presence of noise (Llorà and Goldberg 2003).

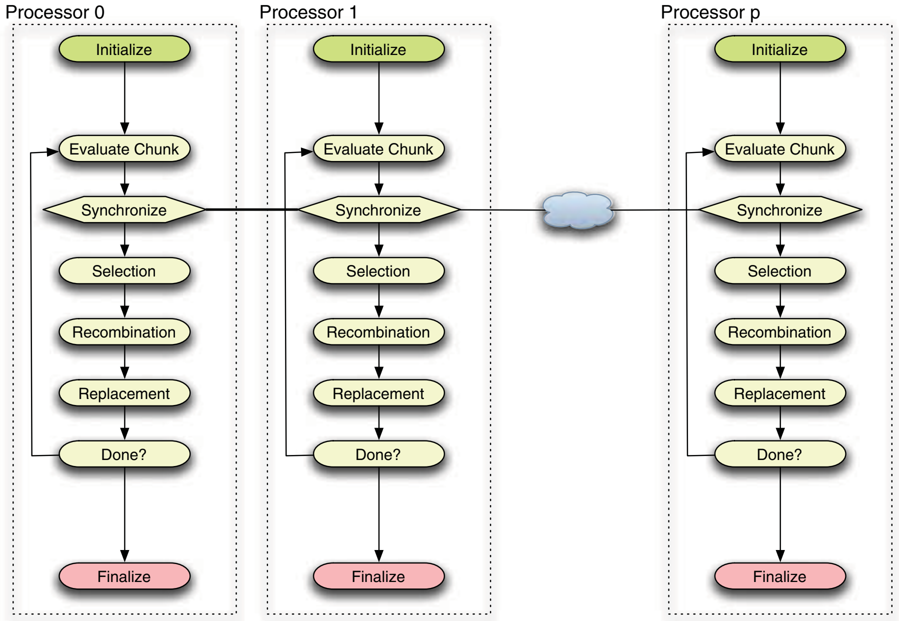


Fig. 6 This figure illustrates the parallel model implemented. Each processor is running the same identical NAX algorithm. They only differ in the portion of the population being evaluated. The population is treated as collection of chunks where each processor evaluates its own assigned chunks sharing the fitness of these individuals with the rest of the processors. This approach minimizes the communication cost

Parallelization may not help much if we need to send large rule sets across the communication network. For such reasons, GBML techniques work very well on moderate complexity problems (Bernadó et al. 2001; Bacardit and Butz 2006). However, they need to be modified to deal with complex and large data set, and also avoid the boundaries imposed by the issues mentioned above.

4.3 NAX: Incremental rule learning for very large data sets

An incremental rule learning approach may alleviate memory footprint requirements by evolving only one rule at a time, hence, reducing the memory requirements. However, one rule by itself cannot solve complex problems. For such a reason, each evolved rule is added to the final rule set, and the covered examples are removed from the current training set. The process is repeated until no instances are left in the training set. This approach already introduced by Cordón et al. (2001) and later also used by Bacardit and Krasnogor (2006) allows maintaining relatively small memory footprints, making feasible processing large data sets—as the prostate tissue classification data set. However, an incremental approach to the construction of the rule set requires paying special attention to the way rules are evolved. For each run of the genetic algorithm used to evolve a rule, we would like to obtain a maximally general and maximally accurate rule, that is, a rule that covers the maximum number of example without making mistakes (Wilson 1995).

```
 1. void evaluate_population ( Population * pp, InstanceSet is, int iDim, int iRows )
 2. {
 3.    int i;
 4.
 5.    /* Compute the fragments of this processor */
 6.    int iFrag = pp->iLen/FCS_processes;
 7.    int iInit = FCS_process_id*iFrag;
 8.    int iLast = (FCS_process_id+1==FCS_processes)?
 9.                      pp->iLen:
10.                      (FCS_process_id+1)*iFrag;
11.    int iCnt  = 0;
12.    int j,k,l;
13.
14.    /* Create the bucket for the broadcast */
15.    float faFit[2*iFrag];
16.    float faTmp[2*iFrag];
17.
18.    /* Evaluate the given chunk assigned to the processor */
19.    for ( i=iInit,iCnt=0 ; i<iLast ; i++,iCnt++ ) {
20.       match_rule_set(pp->prs[i],is,iDim,iRows );
21.       compute_raw_accuracy_fitness_rule_set(pp->prs[i]);
22.       faFit[iCnt] = pp->prs[i]->fFitness;
23.    }
24.
25.    /* Broadcast each of the chunks */
26.    for ( i=0 ; i<FCS_processes ; i++ ) {
27.       MPI_Bcast((i==FCS_process_id)?faFit:faTmp,iCnt,MPI_FLOAT,i,MPI_COMM_WORLD);
28.       if ( i!=FCS_process_id )
29.          for ( l=0,j=i*iFrag, k=(i+1)*iFrag ; j<k ; j++,l++ )
30.             pp->prs[j]->fFitness = faTmp[l];
31.       }
32. }
```

Fig. 7 This figure presents an implementation of the proposed parallel evaluation scheme using `C` and `MPI`. The piece of code presented below is the only one modified to provide such parallelization capabilities. Each processor computes which individuals are assigned to it (lines 6–10), then it computes the fitness (lines 10–23), and then it just broadcast the computed fitness (lines 26–31)

Llorà et al. (2007) have shown that evolving such rules is possible. In order to promote maximally general and maximally accurate rules à la XCS (Wilson 1995), we compute the *accuracy* (α) and the *error* (ε) of a rule (Llorà et al. 2005). The *accuracy* is the proportion of overall examples correctly classified, and the *error* is the proportion of incorrect classifications issued. For simplicity reasons, we use the proportion of correctly issues classifications instead, simplifying the final fitness calculation. Let n_{t+} be the number of positive examples correctly classified, n_{t-} the number of negative examples correctly classified, n_m the number of times a rule has been matched, and n_t the number of examples available. Using these values, the *accuracy* and *error* of a rule r can be computed as:

$$\alpha(r) = \frac{n_{t+}(r) + n_{t-}(r)}{n_t} \tag{3}$$

$$\varepsilon(r) = \frac{n_{t+}(r)}{n_m(r)} \tag{4}$$

Once the *accuracy* and *error* of a rule are known, the fitness can be computed as follows.

$$f(r) = \alpha(r) \cdot \varepsilon(r)^{\gamma} \qquad (5)$$

where γ is the error penalization coefficient. The above fitness measure favors rules with a good classification accuracy and a low error, or maximally general and maximally accurate rules. By increasing γ, we can bias the search towards correct rules. This is an important element because assembling a rule set based on accurate rules guarantees the overall performance of the assembled rule set. In our experiments, we have set γ to 18 to strongly bias the search toward maximally general and maximally accurate rules.

NAX 's efficient implementation of the evolutionary process is based on the techniques described using hardware acceleration—Sect. 4.1.1—and coarse-grain parallelism—Sect. 4.1.2. The genetic algorithm used was a modified version of the *simple genetic algorithm* (Goldberg 1989) using tournament selection ($s = 4$), one point crossover, and mutation based on generating new random boundary elements.

5 Experiments

This section present the results achieved using NAX. To allow the reader to compare with other techniques, we compare the results obtained using NAX on small data sets provided by the UCI repository (Merz and Murphy 1998) to other well-known supervised learning algorithms. Finally, we present the first results on the prostate tissue prediction obtained using NAX. Results focus on the viability of the NAX approach.

5.1 Some UCI repository data sets

The UCI repository (Merz and Murphy 1998) provides several data sets for different machine learning problems. These data sets have been widely used to test traditional machine learning and GBML techniques. Table 1 list the data sets used. Due to the nature of the prostate tissue type classification, we only chose data sets with numeric attributes. Three of these data sets are of relevant interest: (1) son, by far the one with larger dimensionality, (2) gls, the one with large number of classes, (3) tao, proposed by Llorà and Garrell (2001), having complex and non-linear boundaries.

Table 1 Summary of the data sets used in the experiments

ID	Data set	Size	Missing values(%)	Numeric attributes	Nominal attributes	Classes
bre	*Wisconsin Breast Cancer*	699	0.3	9	–	2
bpa	*Bupa Liver Disorders*	345	0.0	6	–	2
gls	*Glass*	214	0.0	9	–	6
h − s	*Heart Stats-Log*	270	0.0	13	–	2
ion	*Ionosphere*	351	0.0	34	–	2
irs	*Iris*	150	0.0	4	–	3
son	*Sonar*	208	0.0	60	–	2
tao	*Tao*	1888	0.0	2	–	2
win	*Wine*	178	0.0	13	–	3

Table 2 Experimental results: percentage of correct classifications and standard deviation from stratified ten-fold cross-validation runs

ID	0–R	C4.5	NAX
bre	65.52 ± 1.16	95.42 ± 1.69	96.43 ± 1.72
bpa	57.97 ± 1.23	65.70 ± 3.84	64.07 ± 8.36
gls	35.51 ± 4.49	65.89 ± 10.47	68.02 ± 8.69
h − s	55.55 ± 0.00	76.30 ± 5.85	75.56 ± 9.39
ion	64.10 ± 1.19	89.74 ± 5.23	89.19 ± 5.27
irs	33.33 ± 0.00	95.33 ± 3.26	94.67 ± 4.98
son	53.37 ± 3.78	71.15 ± 8.54	73.62 ± 9.72
tao	49.79 ± 0.17	95.07 ± 2.11	97.41 ± 0.92
win	39.89 ± 3.22	93.82 ± 2.85	94.34 ± 6.09

Paired *t*-test comparisons showed no statistically significant differences between C4.5 and NAX results

0–R result are just provided as guiding base line

We could have chosen complex algorithms as baselines for NAX . However, we would not be able to use them to repeat the experimentation on the prostate tissue classification domain. The algorithms used in the comparison presented in Table 2 were 0-R (Holte 1993) (a simple base line based on majority class classification) and C4.5 (Quinlan 1993). Results show percentage of correct classifications and standard deviation from stratified ten-fold cross-validation runs. Paired *t*-test comparisons showed no statistically significant differences between the pruned tree produced by C4.5 and NAX results. This experiments also helped validate the distributed implementation proposed by NAX. Further results on empirical comparisons can be found elsewhere (Bernadó et al. 2001; Bacardit and Butz 2006).

5.2 Prostate tissue classification

With the previous results at hand, we ran NAX against the prostate tissue classification data set. The original data set is defined by 93 attributes. In this paper, however, we used the reduced version of this data set proposed by (Fernandez et al. 2005) which contains 20 selected attributes out of the 93 available. The dataset is form by 171,314 records. Our goal was to explore how well NAX could generalize over unseen tissue—this is the first step to be able to address the cancer prediction problem. The other reason that motivated such experimentation was to achieve similar accuracy results as the ones published earlier by Fernandez et al. (2005) using a modified Bayes technique. If NAX could perform at the same level, we will also obtain a set of rules of interest to the spectroscopist. The interpretation of the rules will provide insight on how to interpret the models provided by NAX —which could not be done with the models early used by Fernandez et al. (2005).

We conducted stratified 10-fold cross-validation experiments to measure the generalization capabilities of NAX for this problem. Since the problem was rather small—larger data set are being prepared to be run at the supercomputing facilities provided by the National Center for Supercomputing Applications—we run the ten-fold cross-validation runs in a 3GHz dual core Pentium D computer with 4 GB of RAM. NAX took advantage of the hardware support to speedup the matching process and uses two MPI processes to parallelize—as introduced in Fig. 6—the evaluation of the overall population. Each fold took about one hour to complete, with the entire classification lasting less than half a day. We conducted a simple test of adding a second computer with an identical configuration. The overall time for cross-validation was reduced to half. Rough estimates—which will better measured when larger experiments are conducted on NCSA super computers—show that the sequential portion is around 1:1000 for this small data set. Numbers get better as data set increases, which demonstrates that we will be able to process very large data sets and efficiently exploit larger numbers of processors.

We proposed another measure of effectiveness, namely how many records can be processed per second. Using a single processor with the hardware acceleration mechanisms built into NAX, and the evolved rule set formed by 1,028 rules, the average throughput was around 60,000 records per second. For the prostate tissue classification, it took less than three seconds to classify the entire data set. Once the rule set is learnt, the classification problem falls again into the category of embarrassingly parallel problems (Grama et al. 2003). Since no communication is needed, the speedup grows linearly with the number of processors added—with the proper rule set replication and data set chunking. Thus, with the dual core box used we where able to just double the throughput (120,000 records per second) by chunking the data set and use both processors.

The previous results show the benefits of hardware acceleration and parallelization, but NAX was also able to achieve very competitive classification accuracy in generalization, correctly classifying 97.09 ± 0.09 of the records (pixels) during the stratified ten-fold cross-validation. Figure 8 presents the regenerated prostate tissue classification image presented in Fig. 2 using a rule set assembled by NAX. Figure 8a presents the incorrectly classified pixels. Most of the mistakes by the rule set involve similar tissues with few training records available. This trend was also shown elsewhere (Fernandez et al. 2005). C4.5 does not provide any statistically significant improvement (only a marginal, not statistically significant, 0.7%) and provided large decision trees with more than 5,000 leaves—not to mention the lack of scalability when compared to NAX.

The rule set assembled by NAX represents an incremental assembling of maximally general and maximally accurate rules. Thus, we can compute how the accuracy of such ensemble improves as new rules are added. Figure 9 presents the overall accuracy as rules are added. It shows an interesting behavior for classifying prostate tissue. Using only 20 rules out of the 1,028 evolved ones, the overall accuracy is 90%, the incorrectly classified 1.3% pixels, and 8.7% were left unclassified. After inspecting the misclassified pixels most of them belongs to borders between tissues and mislabeling arises from the image discretization—one pixel containing different tissue types. Table 3 presents the initial four rules that covering 80% of the instances belonging to the two larger tissue types—epithelium and fibrous stroma. Such results are relevant, not only for their accuracy, but also because of the insight they provide to the spectroscopist about the problem structure.

6 Conclusions and further work

This paper has presented the initial results achieved in predicting prostate tissue type using GBML techniques. Being able to classify unseen tissue quickly, reliably, and accurately, is the first step towards the creation of CADx systems that may assist a pathologist diagnosing prostate cancer. We have proposed two main efficiency enhancement techniques for GBML—exploiting hardware parallelization via vector instructions and coarse-grain parallelism via the usage of MPI libraries—which allowed us to approach very large data sets. These techniques, together with an incremental genetics-based rule learning approach to

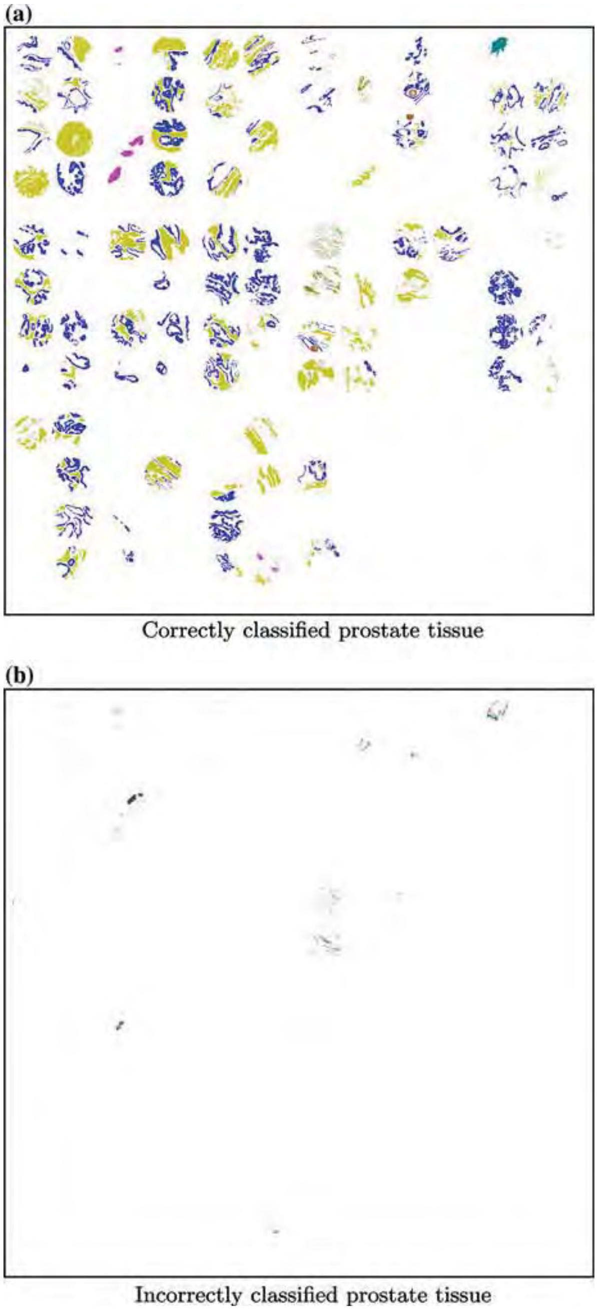

(a)
Correctly classified prostate tissue
(b)
Incorrectly classified prostate tissue

Fig. 8 The figures presented above show the regenerated prostate tissue classification image presented in Fig. 2. (**a**) presents the correctly classified pixels. (**b**) presents the incorrectly classified pixels

assemble rule sets formed by maximally general and maximally accurate rules, have led to the creation of NAX, a system specialized on dealing with large data sets.

Results have shown accurate classification models for prostate tissue along with good scalability of the NAX implementation. Results also reveal peculiarities of the underlying problem structure. With very few rules—20—we were able to correctly classify up to 90%

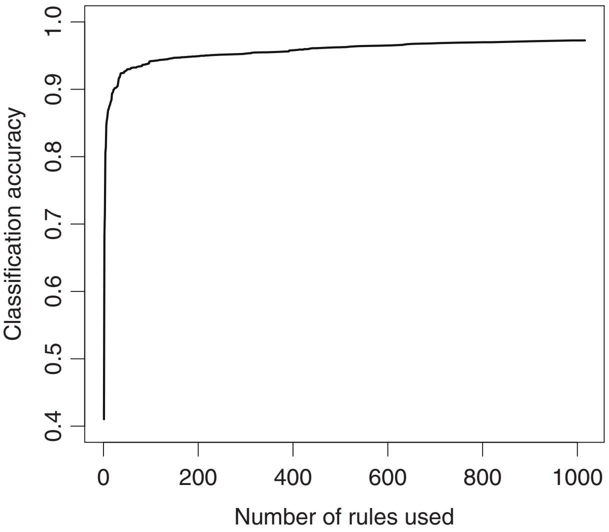


Fig. 9 The rule set as a decision list. The figure presents the classification accuracy as we keep adding rules to the decision list. The first 20 initial rules are able to cover 91% of the records with a classification accuracy of 98.5–90% overall accuracy presented in the figure

Table 3 First top four maximally general and maximally accurate rules that compose the final rule set. The rule set is treated as a decision list, thus we can easily incrementally evaluate the value of the initial four ones

Rule	Rule condition	Tissue type	Accumulated accuracy (%)	Covered records (%)
1.	$0.10 \le a_1 \le 0.25$ ^ $0.00 \le a_4 \le 0.04$ ^ $1.07 \le a_8 \le 2.01$ ^ $-0.07 \le a_{16} \le 0.16$ ^ $0.25 \le a_{17} \le 2.86$ ^ $0.11 \le a_{18} \le 0.21$	→ *Fibrous stroma*	41.32	41.96
2.	$0.03 \le a_1 \le 0.11$ ^ $0.05 \le a_7 \le 0.20$ ^ $1231.88 \le a_{12} \le 1247.90$ ^ $1.98 \le a_{17} \le 3.83$ ^ $0.13 \le a_{18} \le 0.20$	→ *Epithelium*	68.53	69.61
3.	$0.07 \le a_0 \le 0.16$ ^ $0.14 \le a_1 \le 0.41$ ^ $0.71 \le a_{10} \le 1.13$ ^ $1527.54 \le a_{15} \le 1533.80$ ^ $0.65 \le a_{19} \le 1.50$	→ *Fibrous stroma*	71.59	72.75
4.	$0.05 \le a_2 \le 0.09$^ $0.76 \le a_4 \le 1.29$^ $1.80 \le a_6 \le 2.08$^ $0.17 \le a_7 \le 0.24$^ $0.26 \le a_{16} \le 0.53$^ $2.79 \le a_{17} \le 7.01$^ $0.21 \le a_{18} \le 0.32$	→ *Epithelium*	80.78	82.08

of the tissue. Our current work is focused on analyzing the larger data sets containing all the available features and different tissue sources to test the parallelization scalability of NAX on NCSA supercomputers. Once accomplished, the procedure will provide confidence in creating a CADx system to generate a diagnosis based on the evolved models.

Acknowledgments We would like to thank David E. Goldberg for his continual support and encouragement, allowing us to have access to the IlliGAL resources. Thanks also to Kumara Sastry for hallway discussions and to the Automated Learning Group and the Data-Intensive Technologies and Applications at the National Center for Supercomputing Applications for hosting this joint collaboration.This work was sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF, under grant

FA9550-06-1-0370, the National Science Foundation under grant IIS-02-09199, and the National Institute of Health. The US Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation thereon.The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Office of Scientific Research, the National Science Foundation, or the US Government.Rohit Bhargava would like to acknowledge collaborators over the years, especially Dr. Stephen M. Hewitt and Dr. Ira W. Levin of the National Institutes of Health, for numerous useful discussions and guidance. Funding for this work was provided in part by University of Illinois Research Board and by the Department of Defense Prostate Cancer Research Program. This work was also funded in part by the National Center for Supercomputing Applications and the University of Illinois, under the auspices of the NCSA/UIUC faculty fellows program.

References

Amdahl G (1967) Validity of the single processor approach to achieving large-scale computing capabilities. In Proceedings of the American federation of information processing societies conference (AFIPS). 30:483–485 AFIPS

Bacardit J, Butz M (2006) Advances at the frontier of Learning Classifier Systems. Chapter data mining in Learning Classifier Systems: Comparing XCS with GAssist, vol I. Springer

Bacardit J, Krasnogor N (2006) Biohel: Bioinformatics-oriented hierarchical evolutionary learning (Nottingham ePrints). University of Nottingham

Barry A, Drugowitsch J (1997) LCSWeb: the LCS wiki. http://www.lcsweb.cs.bath.ac.uk/

Bernadó E, Llorà X, Garrell J (2001) Advances in Learning Classifier Systems: 4th international workshop (IWLCS 2001). Chapter XCS and GALE: a comparative study of two Learning Classifier Systems with six other learning algorithms on classification tasks. Springer Berlin, Heidelberg, pp 115–132

Bhargava R, Fernandez D, Hewitt S, Levin I (2006) High throughput assessment of cells and tissues: Bayesian classification of spectral metrics from infrared vibrational spectroscopic imaging data. Biochemica et Biophisica Acta 1758(7):830–845

Cantú-Paz E (2000) Efficient and accurate parallel genetic algorithms. Kluwer Academic Publishers

Cordón O, Herrera F, Hoffmann F, Magdalena L (2001) Genetic fuzzy systems. Evolutionary tuning and learning of fuzzy knowledge bases. World Scientific

Fernandez D, Bhargava R, Hewitt S, Levin I (2005) Infrared spectroscopic imaging for histopathologic recognition. Nat Biotechnol 23(4):469–474

Flockhart I (1995) GA-MINER: parallel data mining with hierarchical genetic algorithms (final report). (Technical Report Technical Report EPCCAIKMS-GA-MINER-REPORT 1.0). University of Edinburgh

Gabriel E, Fagg G, Bosilca G, Angskun T, Dongarra J, Squyres J, Sahay V, Kambadur P, Barrett B, Lumsdaine A, Castain R, Daniel D, Graham R, Woodall T (2004) Open MPI: goals, concept, and design of a next generation MPI implementation. In Proceedings of the 11th European PVMMPI Users' group meeting Springer

Goldberg D (1989) Genetic algorithms in search, optimization, and machine learning. Addison-Wesley Professional

Goldberg D (2002) The design of innovation: lessons from and for competent genetic algorithms. Springer

Grama A, Gupta A, Karypis G, Kumar V (2003) Introduction to parallel computing. Addison-Wesley

Holte R (1993) Very simple classification rules perform well on most commonly used datasets. Mach Learn 11:63–91

Lattouf J-B, Saad F (2002) Gleason score on biopsy: is it reliable for predcting the final grade on pathology? BJU Int 90:694–699

Levin I, Bhargava R (2005) Fourier transform infrared vibrational spectroscopic imaging: integrating microscopy and molecular recognition. Annu Rev Phys Chem 56: 429–474

Llorà X (2002) Genetics-based machine learning using fine-grained parallelism for data mining. Doctoral dissertation, Enginyeria i Arquitectura La Salle. Ramon Llull University, Barcelona, Catalonia, European Union

Llorà X (2006) Learning Classifier Systems and other genetics-based machine learning Blog. http://www-illigal.ge.uiuc.edulcs-n-gbml/

Llorà X, Garrell J (2001) Knowledge-independent data mining with fine-grained parallel evolutionary algorithms. In Proceedings of the genetic and evolutionary computation conference (GECCO'2001). Morgan Kaufmann Publishers, pp 461–468

Llorà X, Goldberg D (2003) Bounding the effect of noise in multiobjective Learning Classifier Systems. Evol Comput J 11(3):279–298

Llorà X, Sastry K (2006) Fast rule matching for Learning Classifier Systems via vector instructions. In Proceedings of the 2006 genetic and evolutionary computation conference. ACM Press, pp 1513–1520

Llorà X, Sastry K, Goldberg D (2005) The compact classifier system: motivation, analysis and first results. In Proceedings of the congress on evolutionary computation, vol 1. IEEE press, (Also as IlliGAL TR No 2005019, pp 596–603)

Llorà X, Sastry K, Goldberg D, de la Ossa L (2007) The χ-ary extended compact classifier system: linkage learning in Pittsburgh LCS. In Advances at the frontier of Learning Classifier Systems, vol II. IlliGAL report no 2006015. Springer, pp (in preparation)

Merz CJ, Murphy PM (1998) UCI repository for machine learning data-bases. http://www.ics.uci.edu/~mlearn/MLRepository.html

Mitchell T (1997) Machine learning. McGraw Hill

Orriols-Puig A, Bernadó-Mansilla E (2006) A further look at UCS classifier system. In Proceedings of the 8th annual conference on genetic and evolutionary computation workshop program. ACM Press

Quinlan JR (1993) C4.5: Programs for machine learning. Morgan Kaufmann

Stone C, Bull L (2003) For real! XCS with continuous-valued inputs. Evol Comput J 11(3):279–298

Wilson S (1995) Classifier fitness based on accuracy. Evol Comput 3(2):149–175

Wilson S (2000a) Get real! XCS with continuous-valued inputs. Lect Notes Comput Sci 1813:209–219

Wilson S (2000b) Mining oblique data with xcs. In Revised papers of the 3th international workshop on Learning Classifier Systems (IWLCS 2000). Springer, pp 158–176

Towards Better than Human Capability in Diagnosing Prostate Cancer Using Infrared Spectroscopic Imaging

Xavier Llorà[1], Rohith Reddy[2,3], Brian Matesic[2], and Rohit Bhargava[2,3]

[1]National Center for Super Computing Applications (NCSA)
[2]Department of Bioengineering
[3]Beckman Institute for Advanced Science and Technology
University of Illinois at Urbana-Champaign, Urbana IL 61801
xllora@uiuc.edu, rkreddy2@uiuc.edu, matesic2@uiuc.edu, rxb@uiuc.edu

ABSTRACT

Cancer diagnosis is essentially a human task. Almost universally, the process requires the extraction of tissue (biopsy) and examination of its microstructure by a human. To improve diagnoses based on limited and inconsistent morphologic knowledge, a new approach has recently been proposed that uses molecular spectroscopic imaging to utilize microscopic chemical composition for diagnoses. In contrast to visible imaging, the approach results in very large data sets as each pixel contains the entire molecular vibrational spectroscopy data from all chemical species. Here, we propose data handling and analysis strategies to allow computer-based diagnosis of human prostate cancer by applying a novel genetics-based machine learning technique (NAX). We apply this technique to demonstrate both fast learning and accurate classification that, additionally, scales well with parallelization. Preliminary results demonstrate that this approach can improve current clinical practice in diagnosing prostate cancer.

Categories & Subject Descriptors

I.2.6 [Artificial Intelligence]: Learning–Concept Learning.
I.5.4 [Pattern Recognition]: Applications.
J.3 [Life & Medical Science]: Medical Information Systems.

General Terms

Algorithms, Design, Performance, Experimentation.

Keywords

Genetics-Based Machine Learning, Learning Classifier Systems, Parallelization, Prostate Cancer.

1. INTRODUCTION

Pathologist opinion of structures in stained tissue is the definitive diagnosis for almost all cancers and provides critical input for therapy. In particular, prostate cancer accounts for one-third of noncutaneous cancers diagnosed in US men, and it is a leading cause of cancer-related death. Hence, it is, appropriately, the subject of heightened public awareness and widespread screening. If prostate-specific antigen (PSA) or digital rectal screens are abnormal, a biopsy is considered to detect or rule out cancer. Prostate tissue is extracted, or biopsied, from the patient and examined for structural alterations. The diagnosis procedure involves the removal of cells or tissues, staining them with dyes to provide visual contrast and examination under a microscope by a skilled person (pathologist).

The challenge in prostate cancer research and practice is to provide a novel Due to personnel, tarining, natural variability and biologic differences, the challenge in prostate cancer research and practice is to provide accurate, objective and reproducible decisions. Conventional optical microscopy followed by manual recognition has been demonstrated to be inadequate for this task. [18]. Hence, we have recently proposed developing a practical approach to this problem using chemical, rather than morphologic, imaging. [19]. In this approach, Fourier transform infrared imaging (FTIR) is employed to provide the entire vibrational spectroscopic information from every pixel of a sample's microscopy image. While the first steps of developing novel imaging and sampling technologies is now reliable, [7] the computational challenge of providing robust classification algorithms that can rapidly provide decisions remains. Due to the above advances in imaging and sampling, data from thousands of patients is available to train and validate algorithms for different disease states. While the application and type of data are unique, a further confounding factor required efficiently processing large volumes of data generated by FTIR imaging. The classification problem can be formulated as a supervised learning problem in which several million pixels (hundred of gigabytes) of accurately labeled data are available for model training and validation. The volume of tissue and (future) need for intra-operative diagnoses imply that rapid and accurate diagnoses are crucial to allow physicians to explore all possible courses of action. Under these conditions, traditional supervised learning ap-

GECCO'07, July 7–11, 2007, London, England, United Kingdom.
Copyright 2007 ACM 978-1-59593-697-4/07/0007 ...$5.00.

proaches and implementations do not scale to provide diagnoses in an appropriate time frame. Hence, efficiently processing and learning models from gigabytes of FITR imaging data requires a careful design of the supervised learning algorithm. Moreover, the biological nature of the problem requires that such models be interpretable to provide fundamental new insight into the disease process. Genetics-based machine learning (GBML) techniques take advantage of the *"quasi embarrassing parallelism"* [17] to provide scaleable, fast, accurate, reliable, and interpretable models. In this paper we present an approach engineered to the desired solutiona and constraints of addressing this human task. A modified version of a sequential genetics-based rule learner that exploits massive parallelisms via the message passing interface (MPI) and efficient rule-matching using hardware-oriented operations is developed. We named this system `NAX` [24], and we have shown that its performance is comparable to traditional and genetics-based machine learning techniques on an array of publicly available data sets. We now show that`NAX`—taking advantage of both hardware and software parallelism—is able to provide prostate cancer diagnoses that are human-competitive. In this paper, we present preliminary results supporting this outcome.

The paper is structured as follows. Section 2 provides an overview of our approach towards computer-aided diagnoses for prostate cancer. Procedure and form of the data are summarized in section 3. `NAX` is introduced in section 4, where we describe the basic components and design decisions in this approach. In section 5 we present preliminary results indicating that the approach presented in this paper is human-competitive. Finally, section 6 summarizes some conclusions and further research.

2. PROBLEM DESCRIPTION

Prostate cancer is the most common non-skin malignancy in the western world. The American Cancer Society estimated 234,460 new cases of prostate cancer in 2006 [31]. Recognizing the public health implications of this disease, men are actively screened through digital rectal examinations and/or serum prostate specific antigen (PSA) level testing. If these screening tests are suspicious, prostate tissue is extracted, or biopsied, from the patient and examined for structural alterations. Due to imperfect screening technologies and repeated examinations, it is estimated that more than 1 million people undergo biopsies in the US alone.

2.1 Prostate Cancer Diagnosis

The removal of a small section of prostate is most often accomplished by core biopsy. A needle is inserted into the tissue and several (6-23) samples are obtained from different positions. Biopsy, followed by manual examination under a microscope is the primary means to definitively diagnose prostate cancer as well as most internal cancers in the human body. Pathologists are trained to recognize patterns of disease in the architecture of tissue, local structural morphology and alterations in cell size and shape. Specific patterns of specific cell types distinguish cancerous and non-cancerous tissues. Hence, the primary task of the pathologist examining tissue for cancer is to locate foci of the cell of interest and examine them for alterations indicative of disease.

The specific cells in which cancer arises in the prostate are epithelial cells. While epithelial-origin cancers account for over 85% of all human cancers, they account for more than 95% of prostate cancers. In prostate tissue, epithelial line secretory ducts within the structural cells (collectively termed 'stroma') that allow the tissue to maintain its structure and function. Hence, a pathologist will first locate epithelial cells in a biopsy and, to examine for cancer, will mentally segment them from stroma.

Biopsy samples are prepared in a specific manner to aid in recognition of cells and disease. The sample is sliced thin ($\sim 5\mu m$ thickness), placed on a glass slide and stained with a dye to provide contrast. The most common dye is a mixture of hematoxylin and eosin ($H\&E$), which stains protein-rich regions pink and nucleic acid-rich regions blue. Empty space, lipids and carbohydrates are typically not stained and characterized by white color in images. Staining allows the pathologist to identify cells based on their nucleus and extra-nuclear regions. Patterns of the same cell type characterize structures. For example, epithelial cells arranged in a circular manner around empty space are characteristic of a duct and endothelial cells similarly arranged are characteristic of blood vessels. The empty space enclosed within a duct in pathology images is termed a lumen. The distortion of the circular pattern of epithelial cells around a lumen is characteristic of cancer.

In low severity cancers, lumens are only slightly distorted, while higher grades of cancer display a lack of lumen and simply consist of masses of epithelial cells supported by little stroma. The relative distortion and change in lumen shape is organized into a grading scheme to assess the severity of the disease, Gleason Scoring system, which is the primary measure of disease that defines diagnosis, helps direct therapy and helps predict those at danger of dying from the disease. Since prostate cancer is multi-focal and the disease quite variable, two dominant patterns of epithelial distortion are selected and each is independently graded on a scale of 1-5. The grades are then summed to provide a Gleason score ranging from 2 (low grade cancer) to 10 (maximum danger cancer). This scale has been widely used since its creation in the 1960s and currently forms the clinical standard of practice. Manual Gleason scoring, however, has severe limitations.

2.2 Limitations of Current Practice

Widespread screening for prostate cancer has resulted in a large workload of biopsied men [16], placing an increasing demand on services. Operator fatigue is well-documented and guidelines limit the workload and rate of examination of samples by a single operator (examination speed and throughput). Importantly, inter- and intra-pathologist variation complicates decision-making. The consistency in determining Gleason scores is rather poor. Intra-observer measurements show that a pathologist confirms their own score less than 50% of the time and are ±1 score no more than 80% of cases [2]. Hence, the diagnoses for $\sim 50\%$ of cases may change and may be significantly altered for $\sim 20\%$ of cases ultimately leading to changes in therapy for a patient subset [30]. The numbers are decidedly cause for concern. For example, a recent study including 15 pathologists and 537 prostate cancer patients, 70.8% of Gleason scores were shown to be inaccurate when compared with the patient's final outcome [18]. Second opinions [29] improve assessment and are cost-effective [10], not to mention their utility in mitigating the effects of healthcare costs, lost wages, morbidity, or potential litigation. In summary, the manual recognition of spatial patterns leaves much to be desired from a process perspective and has far-reaching social effects from a public health perspective.

For the reasons underlined above, there is an urgent need for high-throughput, automated and objective pathology tools. We believe that this need is best met by employing the power of computer algorithms and advanced processing to address prostate cancer diagnosis and grading.

The information content of conventionally stained images is limited, inherently non-specific and varies greatly within patient populations and processing conditions. Hence, the information derived from visible microscopy images is fundamentally limited and automated methods of analyzing stained images have failed to provide a sufficiently robust algorithm to diagnose disease. An alternative to morphology-based microscopy are molecular microscopy techniques to probe disease. Molecular technologies for disease diagnosis are an exciting venue for investigations as they promise better diagnostic capabilities through objective means and a multitude of chemicals to provide insight into the changes indicative of the disease process. In particular, spectroscopy tools allow for the measurement of many molecular species simultaneously. Spectroscopic techniques in imaging form, notably using optics, further enable the analysis to be conducted without perturbing the tissue [11]. In this manuscript, we present the analysis of prostate tissue with one such technique, Fourier transform infrared (FTIR) spectroscopic imaging.

2.3 Molecular Imaging

Infrared spectroscopy is a classical technique for measuring the chemical composition of specimens. At specific frequencies, the vibrational modes of molecules are resonant with the frequency of infrared light. By monitoring all frequencies in the region, a pattern of absorption can be created. This pattern, or spectrum, is characteristic of the chemical composition and is hypothesized to contain information that will help determine the cell type and disease state of the tissue. Recently, FTIR spectroscopy has been developed in an imaging sense. Hence, The data are similar to optical microscopy. The first difference is that no external dyes are needed and the contrast in images can be directly obtained from the chemical composition of the tissue. The second is that each pixel in the visible image contains RGB values but in IR imaging contains several thousand values across a bandwidth $(2000 - 14000nm)$ that is ~ 40 times larger than the visible spectrum $(400 - 700nm)$ [7].

3. DATA AND METHODOLOGY

3.1 Experimental Details

Prostate tissues were obtained from Cooperative Human Tissue Network for the tissue array research program (TARP) laboratory. Using these tissues, tissue microarrays were prepared using a Beecher automated tissue arrayer containing a video overlap system and $0.6mm$ needles. Appropriate institutional review board and National Institutes of Health (USA) guidelines for the protection of human subjects were followed. $5\mu m$ sections of tissue were floated on an infrared transmissive optical window for FTIR spectroscopic imaging. Another $5\mu m$ section obtained from the same point on the tissue specimen was observed using traditional microscopy for comparison. Expert pathologists determined the tissue classification using these microscopy samples by staining with $H\&E$. Pathologists' classification were used as the 'gold standard' for comparison with the results from the methods mentioned in this paper.

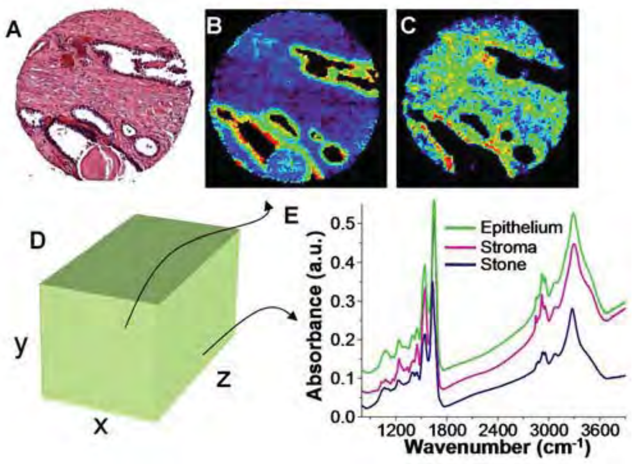


Figure 1: Conventional Staining and Automated Recognition by Chemical Imaging. (A) Typical H&E stained sample, in which structures are deduced from experience by a human. Highlights of specific regions in the manner of H&E is possible using FTIR imaging without stains. (B) Absorption at 1080 cm-1 commonly attributed to nucleic acids and (C) to proteins of the stroma. The data obtained is 3 dimensional (D) from which spectra (E) or images at specific spectral features may be plotted.

Tissues were analyzed using a Michelson interferometer attached to a microscope (Perkin-Elmer Spotlight 300) in transmission mode at a resolution of $4cm^{-1}$ The sample was then raster scanned to obtain images of the entire specimen. Typical specimen size is $600\mu m \times 600\mu m$ with each pixel being $6.25\mu m \times 6.25\mu m$ on the sample plane. Spectra are composed of $1,641$ sample points of the spectral range $4,000 - 720cm^{-1}$. Data acquisition using these techniques required 40 minutes per cylindrical core of the tissue microarray to yield a root mean square signal to noise ratio of $500 : 1$. A typical array was composed of approximately 2.5 million pixels and required 40 GB of storage space.

The data obtained from FTIR imaging is three-dimensional. The $x-$ and $y-$dimensions locate pixels on the tissue-sample plane. The z-dimension values compose the IR spectrum for the corresponding pixel. The spectra can be analyzed to determine what type of tissue (epithelium, stroma, or muscle) the specimen is as well as whether the tissue is malignant or benign. We have developed this technology to provide data from tissue in minutes and employ a high-throughput sampling strategy using Tissue Microarrays (TMA) to obtain data.[19] Samples from multiple tissues, from multiple patients and multiple clinical settings are included in the data set to maximize the sampling of natural variability and ensure the development of robust analysis algorithms. These high-throughput imaging and microarray technologies combine to provide very large data sets—see Figure 1. A typical single core consists of 300×300 pixels on the $x - y$ plane with 1641 bands on the z-axis. A tissue microarray consists of several hundred such cores and analysis of such large datasets (typically, tens of GB) is computationally expensive.

3.2 Data Format

Each pixel's z-dimension contains a spectrum characteristic of the chemical composition of that region of the specimen. Certain spectral quantities provide measures of chemistry. For example, the height of each feature is proportional to its abundance, the peak position is associated with the vibrational identity and peak shape often reflects the multitude of environments around the molecule. Therefore, differences in spectral characteristics can be used in classification and these exact spectral features are termed 'metrics'. For example, the ratio of absorbance of the spectral peak at $1080cm^{-1}$ to the spectral peak at $1545cm^{-1}$ is commonly used to distinguish epithelial from stromal cells. Trained spectroscopists determine these metrics based upon examination of spectral patterns. Hence, the reduction of ull spectra to descriptive metrics forms an intelligent dimensionality reduction strategy. Genetic algorithms form decision rules based upon these metrics to classify pixels by tissue type. Furthermore, the transparency of the genetic algorithms allows the scientist to correlate specific rules to biological features (tissue type and cancer classification) via metrics based upon spectral characteristics.

4. APPROACH

In this section we review related work on the GBML community, highlighting previous efforts to deal with large data sets. We also present the motivation and techniques that lead to the design of `NAX`. Special attention is paid to the description of the hardware and software techniques used, as well as to the design of a scalable GBML algorithm.

4.1 Related Background

Bernadó, Llorà & Garrell [6] presented a first empirical comparison between genetics-based machine learning techniques (GBML) and traditional machine learning approached. The authors reported that GBML techniques were able to perform as well as traditional techniques. Later on, Bacardit & Butz [3] repeated the analysis again obtaining similar results. Most of the experiments presented on both papers were conducted using publicly available data sets provided by the *University of California at Irvine* repository [28]. Most of the data sets are defined over tens of features and up to few thousands of records. However, a key property of GBML approaches is its intrinsic massive parallelism and scalability properties. Cantú-Paz [8] presented how efficient and accurate genetics algorithms could be assembled, and Llorà [21] presented how such algorithms can be efficiently used as machine learning and data mining techniques.

GBML techniques require evaluating candidate solutions against the original data set matching the candidate solutions (e.g. rules, decision trees, prototypes) against all the instances in the data set. Regardless of the GBML flavor used, Llorà & Sastry [25] showed that as the problem grows, the matching process governs the execution time. For small data sets (teens of attributes and few thousands of records) the matching process takes more than 85% of the overall execution time marginalizing the contribution of the other genetic operators. This number easily passes 99% when we move to data sets with few hundreds of attributes and few hundred thousands of records. Such results emphasize one unique facet of GBML approaches: scalability via exploiting massive parallelism. More than 99% of the time required is spent on evaluated candidate solutions. Each solution evaluation is independent of each other and, hence, it can be computed in parallel. Moreover, the evaluation process can also be parallelized further on large data sets by splitting and distributing the data across the computational resources. A detailed description of the parallelization alternatives of GBML techniques can be found elsewhere [21].

Currently available off-the-shelf GBML methods and software distributions [5, 20] do not usually target dealing with very large data sets. Three different works need to be mentioned here. Flockhart [12] proposed and implemented GA-MINER, one of the earliest effort to create data mining systems based on GBML systems that scale across symmetric multi-processors and massively parallel multiprocessors. The work review different encoding and parallelization schemes and conducted proper scalability studies. Llorà [21] explored how fine-grained parallel genetic algorithms could become efficient models for data mining. Theoretical analysis of performance and scalability were developed and validated with proper simulations. Recently, Llorà & Sastry [25] explored how current hardware can be efficiently used to speed up the required matching of solutions against the data set. These three approaches are the basis of the incremental rule learning proposed in the next section to approach very large data sets—such as the prostate tissue classification one.

4.2 The Road to Tractability

`NAX` evolves, one at a time, maximally general and maximally accurate rules. Then, the covered instance are removed and another rule is added to the previously stored one, forming a decision list. This process continues until no uncovered instances are left. Llorà, Sastry & Goldberg [26] showed that maximally general and maximally accurate rules [32] could also be evolved using Pittsburgh-style learning classifier systems. Later, Llorà, Sastry & Goldberg [27] showed that competent genetic algorithms [15] evolve such rules quickly, reliably, and accurately. From these early works, it can be inferred that approaching real-world problems, such as the prostate tissue classification and cancer diagnosis, using GBML techniques may produce the desired byproduct: proper scalability. We discuss next efficient implementation techniques to deal with very large data sets using NAX [24].

4.3 Exploiting the Hardware

Recently, multimedia and scientific applications have pushed CPU manufactures to include support for vector instruction sets again in their processors. Both applications areas require heavy calculations based on vector arithmetic. Simple vector operations such as *add* or *product* are repeated over and over. During 80s and 90s supercomputers, such as Cray machines, were able to issue hardware instructions that took care of basic vector operations. A more constrained scheme, however, has made its way into general-purpose processors thanks to the push of multime-

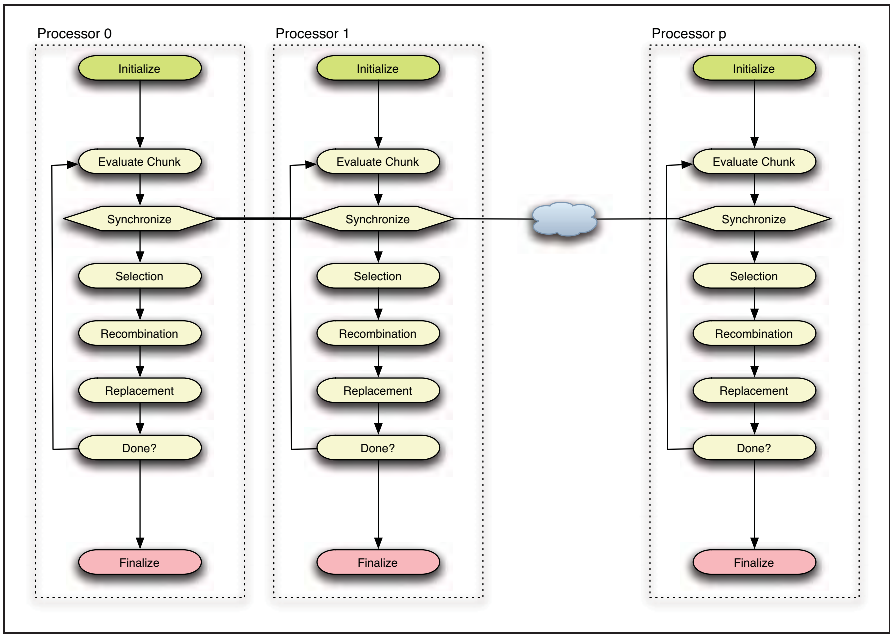


Figure 2: This figure illustrates the parallel model implemented. Each processor is running an identical NAX algorithm. They only differ in the portion of the population being evaluated. The population is treated as collection of chunks where each processor evaluates its own assigned chunk sharing the fitness of these individuals with the rest of processors. This approach minimizes communication cost.

dia and scientific applications. Main chip manufactures—IBM, Intel, and AMD—have introduced vector instruction sets—Altivec, SSE3, and 3DNow$^+$—that allow performing vector operations over packs of 128 bits by hardware. We will focus on a subset of instructions that are able to deal with floating point vectors. This subset of instructions to implemented by hardware vector operations against groups of four floating-point numbers. These instructions are the basis of the fast rule matching mechanism proposed.

Our set of rules seek both to correctly classify the prostate data set and provide biological insight into the rules. All the attributes of the domain are real-value and the conditions of the rules need to be able to express conditions in a $\Re^n$ spaces. We use a rule encoding similar to the one proposed by Wilson [33] and widely used in the GBML community. Rules express the conjunction of tests across attributes. Each test can be defined in multiple fashions, but without loss of generality, we pick a simple interval based one. A simple example of and *if-then* rule, could be expressed as follows:

$$1.0 \leq a_0 \leq 2.3 \wedge \cdots \wedge 10.0 \leq a_n \leq 23 \rightarrow c_1 \qquad (1)$$

Where the condition is the conjunction of the different attribute tests, as introduced earlier, and the condition is the predicting class. We also allow a special condition—`don't care`—which always returns `true` to allow generalized to rules evolve. The rule below illustrates an example of a generalized rule.

$$1.0 \leq a_0 \leq 2.3 \wedge -3.0 \leq a_3 \leq 2 \longrightarrow c_1 \qquad (2)$$

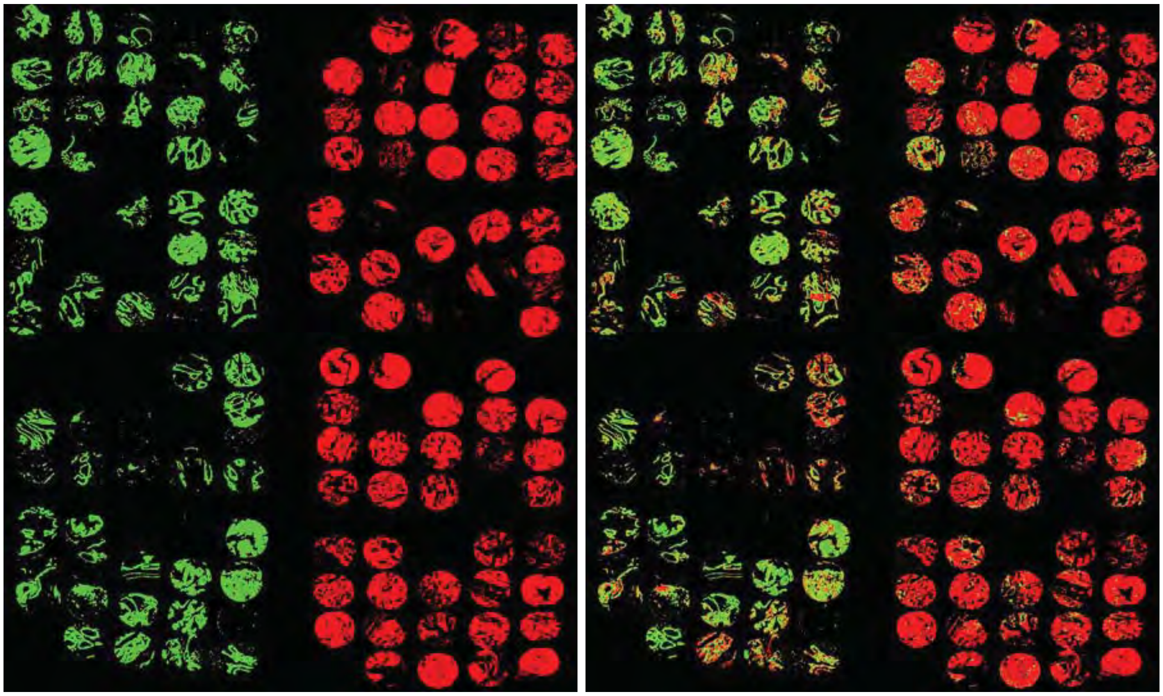
All attributes except a_0 and a_3 were marked as `don't care`.

Matching a rule requires performing the individual tests before the final *and* condition can be computed. Vector instruction sets can help improve the performance of this process by performing four tests at once. Actually, this process can be regarded as four parallel running pipelines. The process can be improved further by stopping the matching process when any one test fails. The code implemented assumes that the two vectors containing the upper and lower bounds are provided and records are stored in a two dimensional matrix. As also shown elsewhere [25], exploiting the hardware available can speed between 3 and 3.5 times the matching process[24].

4.4 Massive Parallelism

Since most of the time is spent on the evaluation of candidate rules when dealing with large data sets, our next goal was to find a parallelization model that could take advantage of this feature. Due to the embarrassing parallelism model [17] for rule evaluation, we designed a coarse-grain parallel model for distributing the evaluation load. Cantú-Paz [8] proposed several schemes, showing the importance of the trade off between computation time and time spent communicating. When designing the parallel model, we focused on minimizing the communication cost. Usually, a feasible solution could be a master/slave one—the computation time is much larger than the communication one. However, GBML approaches tend to use rather large populations, forcing us to send rules to the evaluation slaves and collect the resulting fitness. This scheme also increments sequential instructions that cannot be parallelized, reducing the overall speedup of the parallel implementation as a result of Ambdhals law [1].

To minimize communication cost, each processor runs identical NAX algorithms—all seeded in the same manner, and, hence performing the same genetic operations. They only differ in the portion of the population being evaluated. Thus, the population is treated as collection of chunks where each processor evaluates its own assigned chunk, sharing the fitness of the individuals in its chunk with the rest of processors. in this manner fitness can be encapsulated and broadcasted, maximizing the occupation of the underlying pack-



(a) Original labeled array

(b) Automatically classified array

Figure 3: This figure on the left-hand side presents the original labeled data contained in the P80 array. The figure on the right-hand side presents the reconstructed image based on the predictions issued by the the rule set evolved by NAX. Green represent non cancerous tissue spots; red represent malignant tissue spots.

ing frames used by the network infrastructure. Moreover, this approach also removes the need for sending the actual rules back and forth between processors—as a master/slave approach would require—thus, maintaining the communication to the bare minimum—namely, the fitness. Figure 2 presents a conceptual scheme of the parallel architecture of NAX.

To implement the model presented in Figure 2, we used C and the *open message passing interface* (openMPI) implementation [13]. Each processor computes which individuals are assigned to it. Then it computes the fitness and, finally, it broadcasts the computed fitness. The rest of the process is unchanged. Except for the cooperative evaluation, all the processors generate the same evolutionary trace.

4.5 Lists of Maximally General and Maximally Accurate Rules

One main characteristic of the so-called Pittsburgh-style learning classifier systems—a particular type of GBML—is that the individuals encode a rule set [14, 22, 15]. Thus evolutionary mechanisms directly recombine one rule set against another one. For classification tasks of moderate complexity, the rule sets are not large. For complex problems, however, the potential number of rules required to ensure accurate classification may use prohibitively large amounts of memory. The requirements increase even further in the presence of noise [23]. Hence, this family of GBML techniques works very well on moderate complexity problems [6, 3], but needs to be modified for complex and large data sets.

A sequential rule learning approach may alleviate the requirements by evolving only one rule at a time, hence, reducing the memory requirements [9, 4]. This allows maintaining relatively small memory footprints that makes feasible processing large data sets. However, an incremental approach to the construction of the rule set requires paying special attention to the way rules are evolved. For each run of the genetic algorithm, we would like to obtain a maximally general and maximally accurate rule, that is, a rule that covers the maximum number of examples without making mistakes [32]. NAX (our proposed incremental rule learner) evolves maximally general and maximally accurate rules by computing the *accuracy* (α) and the *error* (ε) of a rule [26]. In a Pittsburgh-style classifier, the *accuracy* may be computed as the proportion of overall examples correctly classified, and the *error* is the proportion of incorrect classifications issued. Once the *accuracy* and *error* of a rule are known, the fitness can be computed as follows.

$$f(r) = \alpha(r) \cdot \varepsilon(r)^{\gamma} \qquad (3)$$

where γ is the error penalization coefficient. We have set γ to 18 to guarantee that the evolutionary process will produce maximally general and maximally accurate solutions. Further details may be found elsewhere [24]. The above fitness measure favors rules with a good classification accuracy and a low error, or maximally general and maximally accurate rules. By increasing γ, we can bias the search towards correct rules. This is an important element because assembling a rule set based on accurate rules guarantees the overall performance of the assembled rule set. NAX's efficient implementation of the evolutionary process is based on the techniques described using hardware acceleration—section

4.3—and coarse-grain parallelism—section 4.4. The genetic algorithm used was a modified version of the *simple genetic algorithm* [14] using tournament selection ($s = 4$), one point crossover, and mutation based on generating new random boundary elements.

5. RESULTS

NAX has shown competitiveness in evolving rule sets that perform as accurately as the ones evolved by other genetics-based machine learning and non-evolutionary machine learning techniques. However, NAXs key element is the ability to deal with large data sets. In this paper, we present preliminary results towards evolving a model capable of correctly classifying pixels as cancerous or non-cancerous. The original array of spots is presented in figure 3(a). Each spot corresponds to a different biopsy sample from a patient. The pixels present in each spot correspond to the epithelial tissue of the biopsy, we supress all other tissue types with a prior classification filter based on Bayesian Likelihood.[7] Each pixel of a spot is defined by 93 different metrics extracted from the processed infrared spectra—as described in section 3. Finally, each pixel in the array was labeled with the diagnostic class provided by a human pathologist. Figure 3(a) presents in green all the non-cancerous pixels while red identifies cancerous ones.

Our goal with the initial experiments here was to demonstrate the usefulness of the proposed approach to computer-aided diagnosis. Our current experimental efforts are planning mass experimentation on several tissue arrays using the Tungsten cluster at the National Center for Supercomputing Applications. These initial experiments were conducted on a dual core Intel Xeon 2.8GHz Linux computer with 1Gb of RAM. NAX was run using both processors. The training time to obtain a model describing all the data took less than ten hours—indicating that very competitive training times can be achieved by just using more processors. The obtained model was able to correctly classify $> 99.99\%$ of the training pixels correctly. However, these results do not illustrate the generalization capabilities of the models evolved by NAX. Hence, we ran a series of ten-fold stratified cross-validation runs [34] to measure generalization and test performance of the evolved models. It is important to mention that tools such as WEKA [34] and other off-the-shelf data miners were not able to handle the volume of data required to evolve a model— either due to the large memory footprint required or by not being able to provide an accurate model in a feasible time period. The results of the cross-validation experiments using NAX correctly classified 87.34% of validation pixels. Such results are more than encouraging, because they show a human-competitive computer-aided diagnosis system is possible. Another interesting property is that a few rules classify a large number of pixels—see Figure 4. Such a result is interesting for the interpretability of the model, since a small number of rules have a great expressiveness, and hence may provide valuable biological insight. Most importantly, they allow us to classify tissue accurately. Subsequent to this pixel level classification, each circular spot in figure 3 was assigned as malignant or benign based on the majority of pixels of he class in the sample. We were able to accurately classify 68 of 69 malignant spots and 70 of 71 benign spots in this manner. While human accuracy is difficult to quantify due to the variation between persons,a generally accepted anecdotal figure is about 5% error rates. The preliminary results we demonstrate here could potentially reduce that five-fold to about 1%, providing a solution to this real-world problem by a combination of novel spectroscopy and advanced machine learning.

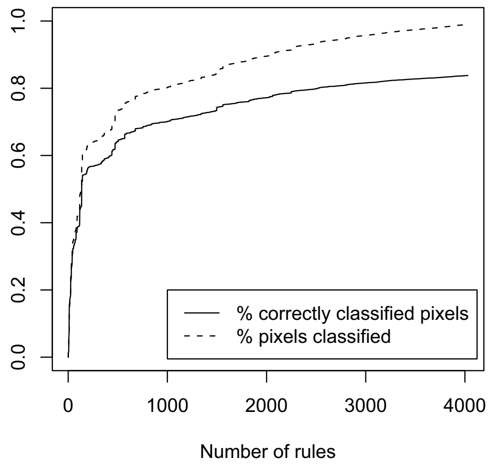


Figure 4: Performance of the evolved model as a function of the number of rules used.

6. CONCLUSION

In this manuscript, we present the application of advanced genetics-based machine learning algorithms to a real-world problem of large scope, namely, the diagnosis of prostate cancer. As opposed to subjective human recognition of disease in tissue using light microscopy, we employed a chemical microscopy approach that required extensive computation but provided a decision without human input. Our development of a learning algorithm based on maximally general and maximally accurate rules was scalable to very large data sets and parallelized to provide learning and classification speed advantages. The algorithm was able to classify a majority of pixels correctly, resulting in overall error rates that were comparable to human examination, the current gold standard of care.

7. ACKNOWLEDGMENTS

This work was sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF, under grant FA9550-06-1-0370, the National Science Foundation under grant ISS-02-09199 at the National Center for Supercomputing Applications, UIUC. The U.S. Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation thereon.

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Office of Scientific Research, the National Science Foundation, or the U.S. Government. We would like to thank Dr. Stephen M. Hewitt, MD PhD, who provided us the samples and expert pathology consultation. The research is also supported by grants to RB from the Research Board, UIUC; the Department of Defense through the Prostate Cancer Research Program and by the National Center for Supercomputing Applications, under the auspices of the NCSA/UIUC faculty fellows program.

8. REFERENCES

[1] G. Amdahl. Validity of the single processor approach to achieving large-scale computing capabilities. In *Proceedings of the American Federation of Information Processing Societies Conference (AFIPS)*, volume 30, pages 483–485. AFIPS, 1967.

[2] M. Amin, D. Grignon, P. Humphrey, and J. Srigley. *Gleason Grading of Prostate Cancer: A Contemporary Approach*. Lippincott Williams & Wilkins: Philadelphia, 2004.

[3] J. Bacardit and M. Butz. *Advances at the frontier of Learning Classifier Systems (Volume I)*, chapter Data Mining in Learning Classifier Systems: Comparing XCS with GAssist, page in press. Springer-Verlag, 2006.

[4] J. Bacardit and N. Krasnogor. Biohel: Bioinformatics-oriented hierarchical evolutionary learning. Nottingham eprints, University of Nottingham, 2006.

[5] A. Barry and J. Drugowitsch. LCSWeb: the LCS wiki, 1997. http://lcsweb.cs.bath.ac.uk/.

[6] E. Bernadó, X. Llorà, and J. Garrell. *Advances in Learning Classifier Systems: 4th International Workshop (IWLCS 2001)*, chapter XCS and GALE: a Comparative Study of Two Learning Classifier Systems with Six Other Learning Algorithms on Classification Tasks, pages 115–132. Springer Berlin / Heidelberg, July 2001.

[7] R. Bhargava, D. Fernandez, S. Hewitt, and I. Levin. High throughput assessment of cells and tissues: Bayesian classification of spectral metrics from infrared vibrational spectroscopic imaging data. *Biochemica et Biophisica Acta*, pages 830–845, 2006.

[8] E. Cantú-Paz. *Efficient and Accurate Parallel Genetic Algorithms*. Kluwer Academic Publishers, 2000.

[9] O. Cordón, F. Herrera, F. Hoffmann, and L. Magdalena. *Genetic Fuzzy Systems. Evolutionary tuning and learning of fuzzy knowledge bases*. World Scientific, 2001.

[10] J. Epstein, P. Walsh, and F. Sanfilippo. Clinical and Cost Impact of Second-opinion Pathology: Review of Prostate Biopsies Prior to Radical Prostatectomy. *American Journal of Surgical Pathology*, 20:851–857, 1996.

[11] D. Fernandez, R. Bhargava, S. Hewitt, and I. Levin. Infrared spectroscopic imaging for histopathologic recognition. *Nature Biotechnology*, 23(4):469–474, 2005.

[12] I. Flockhart. GA-MINER: parallel data mining with hierarchical genetic algorithms (final report). Technical Report Technical Report EPCCAIKMS-GA-MINER-REPORT 1.0, University of Edinburgh, 1995.

[13] E. Gabriel, G. Fagg, G. Bosilca, T. Angskun, J. Dongarra, J. Squyres, V. Sahay, P. Kambadur, B. Barrett, A. Lumsdaine, R. Castain, D. Daniel, R. Graham, and T. Woodall. Open MPI: Goals, concept, and design of a next generation MPI implementation. In *Proceedings of the 11th European PVM/MPI Users' Group Meeting*. Springer, 2004.

[14] D. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Professional., 1989.

[15] D. Goldberg. *The Design of Innovation: Lessons from and for Competent Genetic Algorithms*. Springer, 2002.

[16] S. J. Jacobsen, S. K. Katusic, E. J. Bergstralh, J. E. Oesterling, O. Del, G. G. Klee, C. G. Chute, and M. M. Lieber. Incidence of prostate cancer diagnosis in the eras before and after serum prostate-specific antigen testing. *JAMA*, 274:1445–1449, 1995.

[17] V. Kumar, A. Grama, A. Gupta, and G. Karpis. *Introduction to Parallel Computing: Design and Analysis of Parallel Algorithms*. Benjamin-Cummings Publishing Company, 1994.

[18] J.-B. Lattouf and F. Saad. Gleason score on biopsy: is it reliable for predcting the final grade on pathology? *BJU International*, 90:694–699, 2002.

[19] I. Levin and R. Bhargava. Fourier transform infrared vibrational spectroscopic imaging: integrating microscopy and molecular recognition. *Annual Review of Physical Chemistry*, 56:429–474, 2005.

[20] X. Llorà. Learning Classifier Systems and other Genetics-Based Machine Learning Blog, 2006. http://www-illigal.ge.uiuc.edu/lcs-n-gbml/.

[21] X. Llorà. *Genetics-Based Machine Learning using Fine-grained Parallelism for Data Mining*. PhD thesis, Enginyeria i Arquitectura La Salle. Ramon Llull University, Barcelona, Catalonia, European Union, February, 2002.

[22] X. Llorà and J. Garrell. Knowledge-independent data mining with fine-grained parallel evolutionary algorithms. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO'2001)*, pages 461–468. Morgan Kaufmann Publishers, 2001.

[23] X. Llorà and D. Goldberg. Bounding the effect of noise in multiobjective learning classifier systems. *Evolutionary Computation Journal*, 11(3):279–298, 2003.

[24] X. Llorà, A. Priya, and R. Bhargava. Observer-invariant histopathology using genetics-based machine learning. Technical report, Illinois Genetic Algorithms Laboratory, University of Illinois at Urbana-Champaign (IlliGAL TR No 200627), 2006.

[25] X. Llorà and K. Sastry. Fast rule matching for learning classifier systems via vector instructions. In *Proceedings of the 2006 Genetic and Evolutionary Computation Conference*, pages 1513–1520. ACM Press, 2006.

[26] X. Llorà, K. Sastry, and D. Goldberg. The compact classifier system: Motivation, analysis and first results. In *Proceedings of the Congress on Evolutionary Computation*, volume 1, pages 596–603. IEEE press, 2005. (Also as IlliGAL TR No. 2005019).

[27] X. Llorà, K. Sastry, D. Goldberg, and L. de la Ossa. The χ-ary extended compact classifier system: Linkage learning in Pittsburgh LCS. In *Advances at the frontier of Learning Classifier Systems (Volume II)*, page in preparation. Springer, 2007. IlliGAL report no. 2006015.

[28] C. J. Merz and P. M. Murphy. UCI repository for machine learning data-bases, 1998. http://www.ics.uci.edu/∼mlearn/MLRepository. html.

[29] W. Murphy, I. Rivera-Ramirez, L. Luciani, and Z. Wajsman. Second opinion of anatomical pathology: A complex issue not easily reduced to matters of right and wrong. *J. Urol*, 165:1957–1959, 2001.

[30] J. Nguyen, D. Schultz, A. Renshaw, R. Vollmer, W. Welch, K. Cote, and A. D'Amico. The impact of pathology review on treatment recommendations for patients with adenocarcinoma of the prostate. *Urologic Oncology: Seminars and Original Investigations*, 22:295–299, 2004.

[31] A. C. Society. How Many Men Get Prostate Cancer?, 2006. `http://www.cancer.org/docroot/CRI/content/CRI_2_2_1X_How_many_men_get_prostate_cancer_36.asp?rnav=cri`.

[32] S. Wilson. Classifier fitness based on accuracy. *Evolutionary Computation*, 3(2):149–175, 1995.

[33] S. Wilson. Get real! XCS with continuous-valued inputs. *Lecture Notes in Computer Science*, 1813:209–219, 2000.

[34] I. H. Witten and E. Frank. *Data Mining. Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco, CA., 2000.

Fourier transform infrared spectroscopic imaging: the emerging evolution from a microscopy tool to a cancer imaging modality

Gokulakrishnan Srinivasan and Rohit Bhargava

Department of Bioengineering and Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, Il 61801

INTRODUCTION

The integration of FTIR spectroscopy with microscopy facilitates recording of spatially resolved spectral information, allowing the examination of both the structure and chemical composition of a heterogeneous material. While the first such attempt was over 50 years ago,[1] present day instrumentation largely evolved from the point microscopy detection of interferometric signals that developed in the mid-80s.[2] The successful coupling of interferometry for spectral recording and microscopy for spatial specificity in these systems spurred interest in a variety of fields, including the materials,[3] forensic[4] and biomedical arenas.[5, 6] Point microscopy utilizes an aperture to restrict radiation incident on a sample and permits the recording of spatially localized data. The primary utilities of this form of microscopy lay in acquiring accurate spectra from small-size samples, in determining the chemical structure and composition of heterogeneous phases at specified points and in building a two-dimensional map of the chemical composition of samples. Since the data were acquired at a single point, composition maps could only be acquired by rastering the sample. Hence, the approach was termed mapping or point mapping and involved as many spectral scans as the number of pixels in the map.

The use of focal plane array (FPA) detectors for microscopy[7, 8] allowed for the acquisition of large fields of view in a single interferogram acquisition sweep. The multichannel detection enabled by array detectors was similar to the concept of recording images with charge coupled devices in optical microscopy; hence, the approach was termed imaging. The unique advantages of observing an entire field of view rapidly permitted applications that allowed monitoring of dynamic processes, spatially resolved spectroscopy of large samples or many samples and enhancement of spatial resolution due to retention of radiation throughput that was lost in point microscopy systems due to diffraction at the aperture. Just as for the previous generation of microspectroscopy instruments, applications rapidly followed in the materials[9] and biomedical fields.[10-14] Research activity in this area can be divided into three major categories: instrumentation and sampling methodologies, applications and data extraction algorithms. In this manuscript, we review key advances and recent developments in the context of biomedical imaging. We do not provide comprehensive overview but selectively highlight certain features of importance for cancer-related imaging. Last, we focus on one emerging application area, namely tissue histopathology, and provide illustrative examples from our laboratory indicating the integrative nature of the three in developing protocols.

INSTRUMENTATION, SAMPLING AND DATA HANDLING TECHNIQUES

Instrumentation

Since imaging is largely based on new detectors with unique performance characteristics for spectroscopy, efforts in instrumentation have largely focused on the efficient integration of FPA detectors with interferometers. Due to the size, different electronics and unique noise characteristics of FPAs, an optimization of data acquisition methodology was a primary activity in the initial time period of availability of instrumentation. The first rational attempt at understanding performance and optimizing the data acquisition process revealed the unique noise characteristics that limited the first generation of array detectors.[15] Briefly, this paper established that the general behavior of FTIR spectrometers is generally held for imaging spectrometers but the detector may serve to limit the applicability of established practices in IR spectrometry. An explicit optimization of the data acquisition time revealed several strategies for speeding data collection for both the step scan and rapid scan mode.[16] The first example of rapid-scan FTIR imaging[17]was conducted using asynchronous sampling, followed by descriptions of synchronously triggered sampling and generalized methodologies[18] that could use any detector at any modulation frequency using post-acquisition techniques. Advances in detector technology have now allowed for rapid scan imaging to become routine for large FPA detectors, while innovative new detectors have been developed (first by PerkinElmer) that trade off a large multichannel detection advantage of arrays against the speed of smaller detector arrays to provide a very high performance instrument.[19]

At present, rapid scan imaging has become the mode of choice for most manufacturers and detector sizes have proliferated from the classic 64 x 64 format to range from 16 x 1 to 256 x 256 formats (see figure 1). While the smaller detectors require rastering to image most samples and can provide data of higher quality more efficiently, larger detectors are generally employed for their large field of view and are useful for studying dynamics. It is interesting to note that the linear array approach has an entirely different detector technology and considerations for electronics compared to the two-dimensional FPAs. While it is beyond the scope of this article to discuss the differences, the use of "macro" electronics that are offset from the actual detector and AC mode of operation are the two major differences that affect data. Consequently, comparisons in performance are slightly more complicated. On the large format FPA front, the latest advance seems to be a detector developed jointly by NIH and FBI personnel in 2005. The detector can operate at 16 KHz for 128 x 128 pixel snaps (*Bhargava, Levin, Perlman and Bartick, Unpublished*). This is in the speed regime of single element detectors. Hence, the development can truly lead to the acquisition of an entire image in a single interferometer mirror sweep in the same time that it takes to acquire 1 spectrum with a benchtop IR spectrometer. To handle the large data output, we designed on-chip co-addition and various corrections. We believe that similar detector systems, operating in a fast regime and integrating processing with electronics, are likely to be the technology of tomorrow for FTIR imaging.

The wide variety of instrumentation makes comparisons difficult, especially when manufacturers provide different specifications for instruments. We have proposed a comparison index for these systems based on performance per unit time. Recognizing that spectral resolution, time for scanning, data processing (e.g. apodization) and resultant image size are the primary determinants of performance, a measure can be formulated to describe performance. For a fixed data processing scheme (filtering, apodization etc.), the time taken to acquire 1 megapixels of data for 8 cm^{-1} resolution at a signal to noise ratio (SNR) of 1000:1 is found to be a good measure. We would like to emphasize that the performance is the performance of the entire imaging spectrometer and not due to the detector alone. Efficient coupling of the interferometer and optimization of the optical train will both affect performance as will the correct setup of the experiment. This index also does not consider the ease of use or "user-friendliness" of systems. These are other important considerations and must also be considered by organizations interested in FTIR imaging technology. The issue of time resolution for acquiring data is one such concern. The first approach is the kinetics approach in which the interferometer is repeatedly scanned and imaging data sets are sequentially acquired as quickly as possible. Clearly, rapid scan is favored and the availability of fast readout detectors is mandatory for fast events. The limit to this method is the readout speed of the array (frames in ms) as interferometers can generally be scanned fast enough and the integration time required is typically in the tens of microseconds regime. An example is shown in figure 2 to demonstrate applicability in monitoring polymerization kinetics.

Though rapid scan imaging has displaced the step-scan mode in most new instrumentation, a very important application of the step-scan approach remains in time-resolved imaging.[20-22] Briefly, the method is applicable to systems that can be repeatedly and reproducibly excited and relax back to their ground state. At each mirror retardation, the FPA is repeatedly triggered to acquire data. At the same time, the sample is excited once and the dynamics of excitation and decay of the excited state are monitored. Mirror stepping, data acquisition and sample excitation are all precisely synchronized. Figure 3 demonstrates the synchronization. Time resolved FTIR imaging was first demonstrated using polymer-liquid crystal composites. Examples of the types of data that may be obtained are also shown in figure 3. Last, the technology was extended to provide significantly higher time resolution than could be obtained by the electronics of the detector alone.[23] While FPA detectors are slow compared to single point detectors used in conventional FTIR spectroscopy, the cause is the need to read out data from several thousand pixels and not from the need to record data from all pixels. Hence, by staggering the data recording time over multiple sample excitations, higher temporal resolution may be obtained. With current detectors, a time resolution of ~30 μs should be possible.

Sampling

Interferometer Issues

Among the sampling configurations, the first clearly was the optimization of the microscope for transmission and sampling. Unexpected issues were encountered in initial devices. For example, the detector for the mono-wavelength laser provides a fringe pattern to allow for tracking mirror retardation. The signal from this laser is measured by a small detector located at the center of the beamsplitter (to minimize errors) with an arm that extends out to the edge. When imaged onto the FPA, this laser detector leads to a pattern with low signal levels. Hence, the field of view is not uniform, leading in turn, to lower signal to noise ratios (SNR) for the affected region. Many manufacturers, hence, have re-designed their spectrometers for imaging use. Another manufacturer has avoided this issue by aligning their microscope to sample only the unaffected part of the beam. Since the non-imaging spectrometer did not require imaging and the interferometer was simply coupled to a microscope, these issues were slowly addressed.

Sampling Modes: Transmission, Transmission-reflection, Reflection and Attenuated Total Reflection

A vast majority of studies report the use of transmission sampling. Other major developments have been the incorporation of reflective slides,[24, 25, 26] the integration of ATR elements for both microscopy and large sample imaging, integration of ATR technology with various sample forming accessories, grazing angle accessories and multi-sample accessories. Reflective slides actually result in reflection-absorption that allows the beam to sample the signal twice, though with a different phase and lower signal due to half the objective being used for transmitting light to the sample and the other half being used to acquire light from it. A detailed theoretical understanding of the confounding effects has not been published, though an example of the possible data correction algorithm has been reported. ATR imaging is also highly prevalent and available as attachments to conventional imaging microscopes, using the sample chamber of the spectrometer and using it as a solid immersion lens.[27] We discuss examples of ATR imaging next.

ATR

In the Attenuated Total Reflection (ATR) mode, an IR transmitting crystal of precise geometry of high refractive index is employed as a solid immersion lens. Light is totally reflected at the sample-crystal interface and an evanescent field penetrates into the sample to provide the interaction to be observed using the traveling wave. Since the sample interaction is largely determined by the lens and not by the sample, precise and controlled depth of interaction is available. The sample, however, needs to be in good contact to allow efficient coupling with the evanescent wave. ATR imaging allows users to work with relatively thick sample sections that do not require much sample preparation expertise or time. The first use of ATR imaging was reported by Digilab in analyzing large samples that were not sectioned, as for transmission. ATR imaging microscopy was demonstrated soon after,[28] followed by other novel accessories. There were other unpublished attempts that one of the authors is aware of: In 1999, for example, Snively *et al.* (personal communication, unpublished) demonstrated imaging data from an inverted ZnSe prism acting as a single bounce ATR. Soon after, we employed a Ge crystal but found the signal to noise ratio of the imaging system of that time to be very poor. In addition to the ease of sample preparation, another major advantage of ATR imaging lies in improving the limited spatial resolution of transmission microscopy.[29] The authors assessed that they were able to achieve a spatial resolution of 1μm with a Ge internal reflection element

Both micro and macro sampling has been extensively utilized.[30] A spatial resolution of 3-4 μm using a Ge ATR element was claimed based on more stringent criteria than used previously.[29] Ge, ZnSe and diamond[30] crystals have been the materials of choice for most applications. In particular, Kazarian and co-workers have extensively employed ATR-FTIR imaging for various applications including drug release; polymer/drug formulations and biological systems.[30-33] The same group has provided other innovative sampling configurations for specific experiments, including a compaction cell that allows compaction of a tablet directly on a diamond crystal with a subsequent imaging.[34] The changes in the distribution of a tablet consisting of hydroxypropyl methylcellulose (HPMC) and caffeine upon contact with water were studied. In this manner, conventional dissolution measurements were combined with a concurrent assessment of the compacted tablet structure.[35] As opposed to the organic solvent-polymer dissolution experiments reported earlier, this configuration allows for easy handling and imaging of water-induced dissolution. The setup can also provide high throughput analysis of materials under controlled environments.[36] Microdroplet sample deposition system was combined with a humidity control device to image about 100 samples deposited on the surface of an ATR crystal simultaneously. The approach was extended to 165 samples and were reported to study parallel dissolution of formulations.[37]

Multi-sample Accessories and Sampling

While imaging the structure of materials has been the primary focus of FTIR imaging, a number of applications utilize the imaging of multiple samples. The first examples were from the field of catalyst research.[38] Typically 2-12 samples could be imaged and analyzed under the same conditions. High throughput validation or method development was the primary goal in these studies. Tissue microarrays (TMAs) provide the same function in biomedical imaging. TMAs consist of tens to hundreds of samples arranged on a grid format. This allows for easy visualization of the structure and classification accuracy across many patients and the statistical measures needed for rigorous validation. The primary utility of the multisample image in this case is to provide wide-ranging sampling and convenient archiving or data storage, not necessarily to provide a higher throughput.[14, 39] With the appropriate geometry, many samples can be imaged to understand their dynamics in a concerted fashion. To accommodate the samples, the field of view is often expanded. This results in a lower spatial resolution. For imaging multiple samples, though, the spatial resolution can be conserved but temporal resolution is restricted.

BIOMEDICAL APPLICATIONS

Bone

Bone has been the tissue studied most by FTIR imaging. Bone composition changes with development, environment, genetics, health and disease, is amenable to imaging at the resolution length scale of imaging and has a limited chemical composition that is characterized using IR spectroscopy.[40] For almost 30 years until the late 1980s,[41] bone structure was studied using single element detectors in FTIR spectrometers. Typically, ground bone was analyzed using the conventional KBr pellet method. This pellet method obviously destroyed local structures, precluding an understanding of molecular variations due to disease. Nevertheless, it was sensitive to chemical composition and did provide useful information. With microscopy and now with FTIR imaging, sample integrity is maintained and ability to acquire spectral information at anatomically discrete sites is possible. From the resulting spectra, several important pieces of information can be obtained. For example, a) relative mixture composition of hydroxyapatite and collagen by calculating the ratio of the integrated ν_1, ν_3 phosphate and amide 1 (mineral: matrix ratio), b) carbonate substitution by calculating the ratio of carbonate/phosphate ratio from the ratio of integrated ν_2 carbonate peak (850-900 cm^{-1}) and ν_1, ν_3 phosphate contour (900-1200 cm^{-1}), c) crystallinity of the mineral phase from the ratio of 1030/1020 peak intensity.[42] These assays illustrate several quantities important to bone research and disease diagnoses that can be readily performed. Though a complete discussion is available in the reference[40, 42-44], we pick three illustrative examples demonstrating the applicability in disease and in research.

IR spectral analysis of healthy and disease bone has been reviewed by Boskey *et al*.[42] with particular emphasis on changes in bones composition, physiochemical status of mineral and matrix of bones during osteoporosis and the effect of therapeutics on these parameters. Osteoporosis or porous bone is a bone disease characterized by low bone mass and structural deterioration of bone tissue. This leads to bone fragility and an increased susceptibility to fractures, especially at the hip, spine and wrist. FTIR images of the mineral content and crystallinity in trabecular bone of normal and osteoporotic samples clearly depicts that the trabeculae in diseased tissue are thinner. Moreover, the mineral/matrix ratio in osteoporotic bone is significantly reduced, whereas crystallinity is increased. These advances demonstrate the potential and applicability of the technique to characterize diseased tissue. Bone mineral changes between a healthy mouse model and Fabry diseased (lipid storage disease) mouse model were also analyzed in which globotriaosylceramide (Gb3) accumulates in tissues.[43] No significant differences in the bone mineral properties were observed between Fabry and healthy mice, which might reflect the similar lack of major bone phenotype in human patients with Fabry's disease and may also be related to the developmental age of these animals. The study provides an example of the applicability to laboratory research.

Calcified tissue in biopsies from adults with osteomalica has been studied.[44] Osteomalacia results in a deficiency of the primary mineralization of the matrix, leading to an accumulation of osteoid tissue and reduction in bone's mechanical strength. A decrease in trabecular bone content with absence of changes in matrix or mineral is noticed when iliac crest biopsies of individuals with vitamin D deficient osteomalacia are compared to normal controls. These findings support the assumption that, in osteomalacia, the quality of the organic matrix and of mineral in the centre of the bone does not vary, where as less-than optimal mineralization occurs at the bone surface.

Brain

Monkey brain tissues were one among the first tissues examined by using FTIR imaging.[12] Lately, the applications have experienced a renaissance with applications to the human brain. Grossly, brain can be divided into two types of matter, namely gray matter and white matter. These names derive simply from their appearance to the naked eye. Gray matter consists of cell bodies of nerve cells while white matter consists of the long filaments that extend from the cell bodies - the "telephone wires" of the neuronal network, transmitting the electrical signals that carry the messages between neurons. A visualization of the two compartments formed the first demonstrative application of FTIR microspectroscopic imaging.

FTIR imaging and multivariate statistical analyses (unsupervised hierarchical cluster analysis) were applied alongwith histology and immunohistochemistry in an animal model having Glioblastoma multiform (GBM).[45] GBM is a highly malignant human brain tumor that is considered to be the one of the most difficult to treat effectively.[46] Authors were able to identify the tumor growth as chemically distinct from the surrounding brain tissue. The distribution of the absorbance of amide I in images highlighted high concentrations of proteins in the corpus callosum and regions of basal ganglia for healthy brain. Low absorbance was generally observed in the cortex, whilst a higher absorbance was observed at outer layer of the cortex. For a GBM bearing animal, the highest absorbance was found at the tumor site. In contrast to healthy brain, a lower absorbance of the amide I band was observed at the corpus callosum when compared to that in the cortex and the caudoputamen. The study demonstrates a powerful application of simple analyses that can indicate disease. It also highlights the multitude of spatial and spectral clues that can be use to diagnose or understand the disease.

In addition to primary disease sites, diagnoses metastatic spread from various cancers was also reported.[47] A multivariate classification algorithm was used to distinguish normal tissue from brain metastases successfully and to classify the primary tumor of brain metastases from renal cell carcinoma, lung cancer, colorectal cancer, and breast cancer. In the cluster averaged IR spectra from a brain metastasis of renal cell carcinoma, the main spectral differences were observed for the three tissue regions in the region from 950 to 1200 cm^{-1} and from 1500 to 1700 cm^{-1}. Band intensities of 1026, 1080 and 1153 cm^{-1} are at maximum in the spectrum of black cluster and minimum in the spectrum of light gray cluster. The comparisons of the IR spectra of normal brain tissue and brain metastases of lung, breast cancer and colorectal cancer were made and found that these spectra do not contain spectral features at 1026, 1080 and 1153 cm^{-1} that are indicative of the presence of glycogen. It was concluded that these aforementioned spectral features would be considered as a biomarkers for brain metastases of the primary tumor renal cell carcinoma. In addition to these three bands, the spectral differences were observed for the bands at 1542 and 1655 cm^{-1}, owing to the presence of amide I and amide II vibrations. It is clear from the results that the maximum protein concentrations correlate with minimum glycogen concentrations in the IR image. However, the protein and glycogen properties evident in the IR image are not visible in the unstained cryosection. It is noteworthy that simple univariate analyses provide the end clues to the disease. Even on application of multivariate techniques, the most prominent and easy to understand biomarkers of disease are those defined by conventional spectroscopic knowledge as being important for identification, namely, features and their absorption.

In the cluster–averaged IR spectra of white matter from the three normal brain tissue samples, intense bands at 1060, 1233, 1466, 1735, 2850 and 2920 cm^{-1} due to the high lipid concentration in white matter were noticed. Intensity changes were due to inter-sample and patient to patient variances of the same tissue type. In addition, cluster-averaged IR spectra of a brain metastasis of (renal cell carcinoma, breast cancer, lung cancer, and colorectal cancer) and gray matter of normal brain tissue were compared after baseline subtraction and then normalization with respect to the amide I band. Significant differences in the band positions, intensities and area were observed between these samples which were then used as potential candidates to differentiate normal and tumor tissue and for the identification of the primary tumor. Here, authors used only eight spectroscopic features for LDA model. They were able to classify correctly for three out of three normal brain tissue and 16 out of 17 brain metastases samples. Hence, though univariate analyses and features provide useful recognition, their integration into a multivariate algorithm provides for automated recognition of clinical importance. It may also be argued, however, that it is questionable whether the small numbers of samples employed represent a true performance condition for the algorithm or are simply reflective of bias arising from the clinical setting or sample sources. The advent of faster imaging approaches and advanced sampling techniques like TMAs can allow for larger numbers of samples to be analyzed and such doubts about the validity of studies be put to rest.

Similarly, tissues from rat Glioma models have been characterized and used to discriminate healthy from tumor sections using principal component analysis and K-means.[48] Pseudo color maps reported were constructed on 8-means clusters, where each cluster is consisting of similar spectra. The lipids/protein ratio (1466/1452 cm^{-1}) was found to be decreased and the band at 1740 cm^{-1} became weak and almost vanished as compared to the corresponding bands in the healthy tissue. In addition to the above mentioned differences, significant differences between healthy and tumor affected tissue were observed in the finger print region. In the healthy tissue, a weak band at 1172 cm^{-1}, representing the stretching mode of C-O groups were observed. Reduced intensity as well as shifting of peak to 1190 cm^{-1} was noted for tumor and surrounding tumor spectra. Tumor tissue was observed to contain a decreased intensity of the asymmetric phosphate stretching and C-C stretching and an increased intensity of the symmetric phosphate stretching when compared to the healthy tissue. Variations in lipid features (methylene and methyl stretching) were also observed. The major point here is that the entire spectrum contains numerous points of difference between healthy and diseased tissue. Results were found to be in agreement with those obtained from pathology.[49] The structural difference around the tumor was noted, which could be ascribed to the peritumoral aedoma observed during glioma development. An increase in the permeability of the blood-brain barrier and aggravation in the mass effect of tumors are the rationale for aedoma, which is associated with brain tumor. Fundamental understanding can be enhanced by a complete understanding of the spectral differences but prediction algorithms need only a few measures of the spectral data to be effective.

Breast

Two major applications in breast tissue deal with complications arising from artificial alterations of the tissue and the evolution of cancer. While breast augmentation by implants is highly prevalent, its complications have been discussed more recently. On the other hand, the conventional method for diagnosing and evaluating the prediction of breast disease is a histopathological examination of biopsy samples, a practice that has some shortcomings. For breast implants, a major question is the containment of filling material as its leakage can lead to potential diseases. The silicone gel in implants is very different chemically from surrounding tissue and its presence in tissue sections indicates a definite leak from the implant either due to material failure as a consequence of aging. A spectroscopic image[50] generated from the asymmetric stretching modes of the methyl groups attached to silicon in the gel allowed for the examination of silicone in the tissue. Due to the unique chemical contrast employed in FTIR imaging, such presence can be discerned within the tissue, even when optical microscopy contrast was poor. An example of presence of Dacron (a commercial name for poly(ethylene terepthalate)) fixative patch threads in the breast tissues was shown.[50] It was noted that the technique is capable of rapid analysis within minutes of sectioning the tissue.

A few reports have also applied FTIR imaging for diagnosing breast diseases. Breast tumor tissues were characterized by both FTIR Imaging and point mapping techniques and advantages over the other were evaluated.[51] Similar comparisons had previously been reported for polymeric materials, analyzing both static and dynamic samples.[52] Comparison images from the two methods, imaging data provided a clearer structure in the tumor area than the data obtained from point mapping. Since breast tumor cells are ~10 μm in diameter, point mapping data (with an aperture of 30 μm) would always contains the spectrum of tumor cells as well as from the contributions of other components surrounding the cells. The study clearly indicated that the conventional point mapping approach can fail to detect a small number of malignant cells due to its poor resolution capabilities. Nevertheless, the contamination problem, i.e., the spectral contributions of other components surrounding the cell is found to be less severe in case of ductal carcinoma in situ (DCIS). The study illustrates the need for matching the appropriate level of spatial resolution to the task. While the 30 μm resolution may be appropriate for some applications, it was clearly insufficient for detecting smaller numbers of cells.

Artificial network and K-means cluster analysis have also been employed for the classification of FTIR imaging data from normal and malignant immortalized human breast cell lines.[53] Normal cells, carcinoma cells, mixed normal and carcinoma cells were used. Differences in the spectral backgrounds between the training and test data were observed, which confounds the reproducibility of recorded spectra and, thus, causes the classifier to fail. Using rejection thresholds in the application of the ANN classifier was reported to be helpful in identifying doubtful classifications. Another study[54] reported imaging fibroadenoma, a benign breast tumor. Data were evaluated using unsupervised cluster analysis by utilizing two spectral regions, namely 1000-1500 and 2800-3000 cm^{-1}. The distribution of four main tissue components-epithelium, retro nuclear basal epithelial regions, mantle zone and distant connective tissue were visualized. The spectral features from each component were discussed in detail. Furthermore, comparing epithelia from fibroaedenoma and DCIS, the authors determined that subtle distinctions between the IR characteristics of these two are reproducible. The initial study used tissue from a single patient.

The work was recently extended[55] to diagnose benign and malignant lesions from 22 patients. The study utilized only spectra from well-defined tumor areas owing to the heterogeneity of tissues. Based on the cluster analysis and on comparison with the H & E images, four classes of distinct breast tissue spectra were identified - fibroadenoma (FA), ductal carcinoma in situ (DCIS), connective tissue and adipose tissue. Further, ANNs were developed as an automated classifier to differentiate the four classes. All spectra of connective tissue and adipose tissue were classified correctly, where the spectral features are clearly different from each other and from tumors as well. Differentiating fibroadenoma from DCIS was more difficult. A toplevel/sublevel strategy was further applied and was able to differentiate 93% between fibroadenoma and DCIS spectra by employing principal component analysis. From the mean spectra, it was found that the DCIS has more lipid content than the fibroadenoma. Invasive ductal carcinoma (IDC) could not be well characterized due to contamination from surrounding cells, illustrating the limited spatial resolution.

Cervical Cancer

The cervix is the lower part of the uterus (womb) in which two major types of cancers occur: squamous cell carcinoma and adenocarcinoma. About 80% to 90% of cervical cancers are squamous cell carcinomas, and the remaining 10% to 20% are adenocarcinomas. Less commonly, cervical cancers have features of both squamous cell carcinomas and adenocarcinomas. These are called adenosquamous carcinomas or mixed carcinomas. Typically, the Papanicolaou (Pap) test checks for changes in the exfoliated cells of cervix to find the presence of any infection, abnormal (unhealthy) cervical cells, or cervical cancer. FTIR spectroscopy, micro spectroscopy and FTIR imaging have been widely utilized to study cervical cancer and to perform the same function using computer analyses of spectra.[26, 56-60] While the first reports in diagnosing cervical cancer are now generally not regarded as leading to solutions,[56] two groups have provided definitive proof of the potential of IR spectroscopy by careful microscopy studies.[26, 57, 45, 59, 60] While FTIR images of the amide I and υ_{asy} PO_2^- bands with H&E stained image were compared and only a rough correlation with the pathological features or cell types were obtained, cluster maps of two, five and eight clusters resulting from UHC analysis for the whole spectrum demonstrated good segmentation. In five clusters, most cell types are apparent including superficial (1), intermediate (2), parabasal (3), and connective tissue (5) upon correlation with the stained image. As in univariate images, the connective tissue region (5) is split in to two clusters. Furthermore, by comparing between the UHC analysis of the whole spectrum and only the amide I region, authors demonstrated that minimizing the spectral region for analysis and using fewer clusters does not lead to the loss of useful information. Both univariate FTIR and multivariate images of the sample with several endocervical ducts within the connective tissue were shown. These endocervical ducts lined with columnar endocervical cells were apparent in all those images, in particular even with two clusters.

Cultures derived from cervical cancer cells (HeLa) are one of the most popular model systems and have been studied using FTIR imaging.[61] The cells were directly grown as sparse monolayers onto low-e slides. FTIR image of amide I band region was shown; where large differences in spectral intensities associated with the cells were observed even though these cells are from a homogeneous and exponential cell culture. Cluster analyses of normalized spectra shows distinct differences that were not appreciated in the univariate image. Similarly,[62] IR imaging with fuzzy C- means clustering and hierarchical cluster analysis were utilized to study the thin sections of cervix uteri encompassing normal, precancerous and squamous cell carcinoma. These studies demonstrate that IR imaging, in combination with multivariate techniques, is capable of segmenting cervical tissues in a manner that is comparable to H&E stained image differentiation and is significantly more sensitive in terms of the chemical composition of the cells – whether it be due to metabolic or disease reasons.

Prostate

Prostate cancer is the most prevalent internal cancer in the US.[63] Hence, its pathologic diagnosis and correct interpretation of disease state is crucial.[64] FTIR imaging has been proposed as solution that can potentially help pathologists by providing an objective and reproducible assessment of disease in a manner that is easily understood by clinicians. It is also a good model system for the development of FTIR imaging protocols. We first review progress in the field and then describe efforts in our and collaborator's laboratories towards formulating a practical algorithm for prostate cancer pathology. While a number of studies examined human prostate tissue with IR spectroscopy[65-68] microscopy approaches have recently been extensively utilized to study both fundamental properties of prostate tissue and to determine structural units in normal and disease states.[69-75] An understanding of the tissue is now emerging as a result of these studies. While the fundamental properties of the tissue are being examined, we have focused on developed statistically validated diagnostic methods.

We have utilized high throughout imaging with the express purpose of correlating spectra to clinical practice.[39, 64, 76] It is instructive to first examine the approaches of some previous studies and then describe our approach in some detail. A variety of techniques have been reported for analyzing prostate tissue, including unsupervised multivariate data analysis techniques such as agglomerative hierarchical clustering (AH), fuzzy C-means (FCM), or k-means (KM) clustering to construct infrared spectral maps of tissue structures.[77] The results from these multivariate techniques confirmed the standard histopathological techniques and found out to be helpful for identifying and discriminating the tissues structures. Agglomerative hierarchical clustering was found to be the best method among the cluster imaging methods in terms of segmenting the tissue. While these techniques comprise one end of the approach in using large spectral regions and completely objective methods, the other extreme has also proven to be useful. In the second paradigm, careful examination of the spectral data yields some measures that prove useful. For example, the ratio of peak areas at 1030 and 1080 cm^{-1}, corresponding to the glycogen and phosphate vibrations respectively were utilized as a diagnostic marker for the differentiation of benign from malignant cells.[69] Authors summarized that the use of this ratio in association with FTIR spectral imaging provides a basis for estimating areas of malignant tissue within defined regions of a specimen. While it may be argued that the former is not based on clinical knowledge and is more suited for discovery, it also involves the choice of selecting specific number of clusters and their subsequent interpretation. The latter is based on a single parameter whose utility for universal diagnoses remains to be tested. Nevertheless, these studies indicate that both approaches provide information about the tissue that is useful.

Our approach has used elements from both pattern recognition and spectroscopic analyses of univariate measures. [39, 76] In all cases, one starts with the acquired imaging data (figure 4). Since the data set is large (typically 10-1000 GB), it is advisable to reduce the dimensionality of data using some numerical procedure. Compression algorithms, principal components analyses or simply storing only the information needed for classification (if the algorithm is known) is useful. We sought expressly to relate the recorded IR imaging data to clinical knowledge base. Hence we started with a model that is derived from clinical practice. Clearly, the approach limits the discovery of new knowledge but it assures the clinician that all quantities of importance for diagnoses will be considered. The acquired data is labeled with known cell identity or disease states. These pixels are best identified by a combination of very careful manual labeling and test for absorbance fidelity.[78] Spectra from the label regions are employed via average values, medians and standard deviation analyses to determine a set of spectral features that are descriptive of the major features of all spectra. We first note that the characteristic IR absorbance spectra of ten histological classes comprising prostate tissue look similar. Though small differences in spectral features were observed at many frequencies, summary statistics are limited in their examination of spectra for classification. Further, the small differences indicate that noise and biological variability may render univariate measures less reliable. The large number of classes usually implies that univariate analyses cannot distinguish all histological classes present in the tissues and hence the need for multivariate analyses is apparent. Here the similarity of the spectral features for all classes works in our favor. Very similar baseline points are obtained from an analysis of all spectra and only subtle feature differences are noted to distinguish the various class spectra. Hence, unknown spectra can be processed in the same specified manner, without introducing any bias. Each of these features is termed a metric to denote that it is a useful measure of the spectrum. Individual metrics can allow segmentation of various tissue types if they are sufficiently different in a sampled population.

We then employ the equivalent of a t-test in that the overlap between the absorbance distributions of metrics is determined and equated to the error in prediction. The metrics are arranged in the order of increasing overlap. Hence, we have an ordered set that differentiates at least two classes. To obtain overall accuracy, we employ a modified Bayesian algorithm to provide the probability of each class for every pixel. This fuzzy result is employed to determine the area under the curve (AUC) of a receiver operating characteristic (ROC) curve. The ROC curve is built from accepting the probability of each class at an increasing threshold that varies between 0 and 1. For optimized threshold values, the fuzzy classification is turned into a classified image, where each pixel is assigned a distinct class. We note that the method incorporates analysis of all spectral features, a selection of the best features based on statistical analysis of data and an optimal prediction of the class of each pixel based on an objective selection rule from the fuzzy classification. The method is very powerful in that it employs spectral features that are ordinarily employed by spectroscopists as metrics, which permits a spectroscopic analysis of the basis of decision-making. Further, the method explicitly obtains the fuzzy rule data for final classification. The value of the rule data for each class is actually the probability of belonging to the class without consideration for the prior prevalence of the class. Hence, the method can allow direct comparisons between performances for different classes. The dependence of the process on various experimental parameters has also been reported.

The complication inherent in translating the results from small data set of patients to clinical applications is well recognized in the spectroscopy community. The variability in data, arising from variations within and between patients, sample preparation and handling, is likely to provide noisy estimates of performance. Hence, statistical stability may be obtained by examining a large number of samples. Similarly, large number of patients may be employed to provide calibration models, likely improving the robustness of the developed algorithm. We have described a high throughput sampling method from tissues.[14, 39, 76] Briefly, the approach uses a combinatorial sampling of tissue type and pathology to first acquire small sections of tissues from large archival cases. These small sections are arranged in a grid pattern and placed on the same substrate. The sample is termed a tissue microarray to reflect the similarity with cDNA microarrays. For spectroscopic imaging and the development of automated algorithms, the approach represents a large number of cases that can be used both for accurate prediction algorithm building and for extensive validations. The same approach is likely to prove useful for extensions to determining pathology. Figure 5 demonstrates the typical workflow of a validation algorithm and methods used for statistical comparison. We strongly suggest a variety of methods for measuring performance as each method has its own advantages and disadvantages. For example, summary measures from ROC curves only provide information about accuracy but do not provide which class the inaccuracies arise from. Similarly, confusion matrices provide cross-class information but do not provide global performance measures in the mold of ROC curves.

OUTLOOK

FTIR imaging has experienced rapid growth in the past 10 years and is increasingly being applied to biomedical tissue, especially for the analyses of cancer. The major trends emerging in instrumentation include faster detectors and novel modes of data collection (e.g. time –resolved imaging), of sampling (e.g. ATR) and application areas. For biomedical samples, the information content is quite rich and is often available through simple univariate analysis. For more complex applications, e.g. cancer diagnoses, the data acquisition, sampling and data analyses must be integrated in a coherent manner to provide a practical solution. We anticipate that the technology and its application to biomedical problem will continue to grow with the cooperation of instrument manufacturers, applications scientists, numerical methods developers and communities that can utilize the information effectively, e.g. pathologists or surgeons.

References

1.Gore, R. C.; Barnes, R. B.; Petersen, E., Infrared Absorption of Aqueous Solutions of Organic Acids and Their Salts. *Anal. Chem.* **1949,** 21, (3), 382-386.

2.Kwiatkoski, J. M.; Reffner, J. A., FT-IR microspectrometry advances. *Nature* **1987,** 328, (6133), 837-838.

3.Koenig, J. L., *Microspectroscopic imaging of polymers*. American Chemical Society Washington, DC: 1998.

4.Williams, D. K.; Schwartz, R. L.; Bartick, E. G., Analysis of latent fingerprint deposits by infrared microspectroscopy. *Appl Spectrosc* **2004,** 58, (3), 313-6.

5.Petrich, W., Mid-Infrared and Raman Spectroscopy for Medical Diagnostics *Applied Spectroscopy Reviews* **2001,** 36, (2), 181-237.

6.Naumann, D., FT-infrared and FT-Raman Spectroscopic in Biomedical Research *Applied Spectroscopy Reviews* **2001,** 36, (2), 239-298.

7.Lewis, E. N.; Levin, I. W., Vibrational Spectroscopic Microscopy: Raman, Near-Infrared and Mid-Infrared Imaging Techniques. *Microscopy and Microanalysis* **1995,** 1, (01), 35-46.

8.Lewis, E. N.; Treado, P. J.; Reeder, R. C.; Story, G. M.; Dowrey, A. E.; Marcott, C.; Levin, I. W., Fourier transform spectroscopic imaging using an infrared focal-plane array detector. *Anal Chem* **1995,** 67, (19), 3377-81.

9.Bhargava, R.; Wang, S. Q.; Koenig, J. L., FTIR microspectroscopy of polymeric systems. *Advances in Polymer Science* **2003,** 163, 137-191.

10.Mendelsohn, R.; Flach, C. R.; Moore, D. J., Determination of molecular conformation and permeation in skin via IR spectroscopy, microscopy, and imaging. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2006,** 1758, (7), 923-933.

11.Garidel, P.; Boese, M., Mid infrared microspectroscopic mapping and imaging: A bio-analytical tool for spatially and chemically resolved tissue characterization and evaluationof drug permeation within tissues. *Microsc Res Tech* **2007,** 70, (4), 336-349.

12.Lewis, E. N.; Gorbach, A. M.; Marcott, C.; Levin, I. W., High-fidelity Fourier transform infrared spectroscopic imaging of primate brain tissue. *Applied Spectroscopy* **1996,** 50, (2), 263-269.

13.Lewis, E. N.; Kidder, L. H.; Levin, I. W.; Kalasinsky, V. F.; Hanig, J. P.; Lester, D. S., Applications of fourier transform infrared imaging microscopy in neurotoxicity. *Ann N Y Acad Sci* **1997,** 820, 234-46; discussion 246-7.

14.Levin, I. W.; Bhargava, R., Fourier transform infrared vibrational spectroscopic imaging: integrating microscopy and molecular recognition. *Annu Rev Phys Chem* **2005,** 56, 429-74.

15.Snively, C. M.; Koenig, J. L., Characterizing the Performance of a Fast FT-IR Imaging Spectrometer. *Applied Spectroscopy* **1999,** 53, (2), 170-177.

16.Bhargava, R.; Levin, I. W., Fourier transform infrared imaging: theory and practice. *Anal. Chem* **2001,** 73, (21), 5157-5167.

17.Snively, C. M.; Katzenberger, S.; Oskarsdottir, G.; Lauterbach, J., Fourier-transform infrared imaging using a rapid-scan spectrometer. *Opt. Lett* **1999,** 24, 1841-1843.

18.Huffman, S. W.; Bhargava, R.; Levin, I. W., Generalized Implementation of Rapid-Scan Fourier Transform Infrared Spectroscopic Imaging. *Applied Spectroscopy* **2002,** 56, (8), 965-969.

19. *Spectrochemical Analysis using Infrared Multichannel Detectors, Bhargava, R.*

Levin, I. W. (Eds). Blackwell Publishing, Sheffield, UK: 2005.
20.Bhargava, R.; Levin, I. W., Time-resolved Fourier transform infrared spectroscopic imaging. *Appl Spectrosc* **2003,** 57, (4), 357-66.
21.Bhargava, R.; Levin, I. W., Noninvasive Imaging of Molecular Dynamics in Heterogeneous Materials. *Macromolecules* **2003,** 36, (1), 92-96.
22.Bhargava, R.; Levin, I. W., Gram-Schmidt orthogonalization for rapid reconstructions of Fourier transform infrared spectroscopic imaging data. *Appl Spectrosc* **2004,** 58, (8), 995-1000.
23.Bhargava, R.; Levin, I. W., Enhanced Time-Resolved Fourier Transform Infrared Spectroscopic Imaging for Reversible Dynamics. *J. Phys. Chem. A* **2004,** 108, (18), 3896-3901.
24.O'Leary, T. J.; Engler, W. F.; Ventre, K. M., Infrared Microspectroscopy of Human Tissue. *Applied Spectroscopy* **1989,** 43, (6), 1095-1097.
25.Marcott, C.; Story, G. M.; Dukor, R. K., Infrared Spectral Imaging of H&E-Stained Breast Tissue Biopsies. *Microscopy and Microanalysis* **2004,** 10, (S02), 182-183.
26.Romeo, M.; Mohlenhoff, B.; Jennings, M.; Diem, M., Infrared micro-spectroscopic studies of epithelial cells. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2006,** 1758, (7), 915-922.
27.Patterson, B. M.; Havrilla, G. J., Attenuated Total Internal Reflection Infrared Microspectroscopic Imaging Using a Large-Radius Germanium Internal Reflection Element and a Linear Array Detector. *Applied Spectroscopy* **2006,** 60, (11), 1256-1266.
28.Sommer, A. J.; Tisinger, L. G.; Marcott, C.; Story, G. M., Attenuated Total Internal Reflection Infrared Mapping Microspectroscopy Using an Imaging Microscope. *Applied Spectroscopy* **2001,** 55, (3), 252-256.
29.Otts, D. B.; Zhang, P.; Urban, M. W., High Fidelity Surface Chemical Imaging at 1000 nm Levels: Internal Reflection IR Imaging (IRIRI) Approach. *Langmuir* **2002,** 18, (17), 6473-6477.
30.Chan, K. L. A.; Kazarian, S. G., New opportunities in micro- and macro-attenuated total reflection infrared spectroscopic imaging: Spatial resolution and sampling versatility. *Applied Spectroscopy* **2003,** 57, (4), 381-389.
31.Chan, K. L.; Hammond, S. V.; Kazarian, S. G., Applications of attenuated total reflection infrared spectroscopic imaging to pharmaceutical formulations. *Anal Chem* **2003,** 75, (9), 2140-6.
32.Colley, C. S.; Kazarian, S. G.; Weinberg, P. D.; Lever, M. J., Spectroscopic imaging of arteries and atherosclerotic plaques. *Biopolymers* **2004,** 74, (4), 328-35.
33.Kazarian, S. G.; Chan, K. L. A., Applications of ATR-FTIR spectroscopic imaging to biomedical samples. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2006,** 1758, (7), 858-867.
34.van der Weerd, J.; Andrew Chan, K. L.; Kazarian, S. G., An innovative design of compaction cell for in situ FT-IR imaging of tablet dissolution. *Vibrational Spectroscopy* **2004,** 35, (1-2), 9-13.
35.van der Weerd, J.; Kazarian, S. G., Combined approach of FTIR imaging and conventional dissolution tests applied to drug release. *J Control Release* **2004,** 98, (2), 295-305.
36.Chan, K. L.; Kazarian, S. G., Fourier transform infrared imaging for high-throughput analysis of pharmaceutical formulations. *J Comb Chem* **2005,** 7, (2), 185-9.
37.Chan, K. L.; Kazarian, S. G., ATR-FTIR spectroscopic imaging with expanded field of view to study formulations and dissolution. *Lab Chip* **2006,** 6, (7), 864-70.
38.Snively, C. M.; Oskarsdottir, G.; Lauterbach, J., Chemically sensitive parallel analysis of combinatorial catalyst libraries. *Catalysis Today* **2001,** 67, (4), 357-368.

39.Fernandez, D. C.; Bhargava, R.; Hewitt, S. M.; Levin, I. W., Infrared spectroscopic imaging for histopathologic recognition. *Nat Biotechnol* **2005,** 23, (4), 469-74.
40.Boskey, A. L.; Mendelsohn, R., Infrared spectroscopic characterization of mineralized tissues. *Vibrational Spectroscopy* **2005,** 38, (1-2), 107-114.
41.Posner, A. S.; Duyckaerts, G., Infrared study of the carbonate in bone, teeth and francolite. *Experientia* **1954,** 10, (10), 424-5.
42.Boskey, A.; Mendelsohn, R., Infrared analysis of bone in health and disease. *J Biomed Opt* **2005,** 10, (3), 031102.
43.Boskey, A. L.; Goldberg, M.; Kulkarni, A.; Gomez, S., Infrared imaging microscopy of bone: Illustrations from a mouse model of Fabry disease. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2006,** 1758, (7), 942-947.
44.Faibish, D.; Gomes, A.; Boivin, G.; Binderman, I.; Boskey, A., Infrared imaging of calcified tissue in bone biopsies from adults with osteomalacia. *Bone* **2005,** 36, (1), 6-12.
45.Bambery, K. R.; Schultke, E.; Wood, B. R.; MacDonald, S. T. R.; Ataelmannan, K.; Griebel, R. W.; Juurlink, B. H. J.; McNaughton, D., A Fourier transform infrared microspectroscopic imaging investigation into an animal model exhibiting glioblastoma multiforme. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2006,** 1758, (7), 900-907.
46. American Brian Tumor Association.
47.Krafft, C.; Shapoval, L.; Sobottka, S. B.; Schackert, G.; Salzer, R., Identification of primary tumors of brain metastases by infrared spectroscopic imaging and linear discriminant analysis. *Technol Cancer Res Treat* **2006,** 5, (3), 291-8.
48.Amharref, N.; Beljebbar, A.; Dukic, S.; Venteo, L.; Schneider, L.; Pluot, M.; Vistelle, R.; Manfait, M., Brain tissue characterisation by infrared imaging in a rat glioma model. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2006,** 1758, (7), 892-899.
49.Kneipp, J.; Lasch, P.; Baldauf, E.; Beekes, M.; Naumann, D., Detection of pathological molecular alterations in scrapie-infected hamster brain by Fourier transform infrared (FT-IR) spectroscopy. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **2000,** 1501, (2-3), 189-199.
50.Kidder, L. H.; Kalasinsky, V. F.; Luke, J. L.; Levin, I. W.; Lewis, E. N., Visualization of silicone gel in human breast tissue using new infrared imaging spectroscopy. *Nat Med* **1997,** 3, (2), 235-237.
51.Fabian, H.; Lasch, P.; Boese, M.; Haensch, W., Mid-IR microspectroscopic imaging of breast tumor tissue sections. *Biopolymers - Biospectroscopy Section* **2002,** 67, (4-5), 354-357.
52.Bhargava, R.; Wall, B. G.; Koenig, J. L., Comparison of the FT-IR mapping and imaging techniques applied to polymeric systems. *Applied Spectroscopy*
2000, 54, 470-479.
53.Zhang, L.; Small, G. W.; Haka, A. S.; Kidder, L. H.; Lewis, E. N., Classification of Fourier transform infrared microscopic imaging data of human breast cells by cluster analysis and artificial neural networks. *Appl Spectrosc* **2003,** 57, (1), 14-22.
54.Fabian, H.; Lasch, P.; Boese, M.; Haensch, W., Infrared microspectroscopic imaging of benign breast tumor tissue sections. *Journal of Molecular Structure* **2003,** 661-662, (1-3), 411-417.
55.Fabian, H.; Thi, N. A. N.; Eiden, M.; Lasch, P.; Schmitt, J.; Naumann, D., Diagnosing benign and malignant lesions in breast tissue sections by using IR-microspectroscopy. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2006,** 1758, (7), 874-882.

56.Wong, P. T.; Wong, R. K.; Caputo, T. A.; Godwin, T. A.; Rigas, B., Infrared spectroscopy of exfoliated human cervical cells: evidence of extensive structural changes during carcinogenesis. *Proc Natl Acad Sci U S A* **1991,** 88, (24), 10988-92.

57.Boydston-White, S.; Romeo, M.; Chernenko, T.; Regina, A.; Miljkovic, M.; Diem, M., Cell-cycle-dependent variations in FTIR micro-spectra of single proliferating HeLa cells: Principal component and artificial neural network analysis. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2006,** 1758, (7), 908-914.

58.Walsh, M. J.; German, M. J.; Singh, M.; Pollock, H. M.; Hammiche, A.; Kyrgiou, M.; Stringfellow, H. F.; Paraskevaidis, E.; Martin-Hirsch, P. L.; Martin, F. L., IR microspectroscopy: potential applications in cervical cancer screening. *Cancer Letters* **2007,** 246, (1-2), 1-11.

59.Bambery, K. R.; Wood, B. R.; Quinn, M. A.; McNaughton, D., Fourier transform infrared imaging and unsupervised hierarchical clustering applied to cervical biopsies. *Australian Journal of Chemistry* **2004,** 57, (12), 1139-1143.

60.Wood, B. R.; Chiriboga, L.; Yee, H.; Quinn, M. A.; McNaughton, D.; Diem, M., Fourier transform infrared (FTIR) spectral mapping of the cervical transformation zone, and dysplastic squamous epithelium. *Gynecol Oncol* **2004,** 93, (1), 59-68.

61.Diem, M.; Romeo, M. J.; Boydston-White, S.; Matthaus, C., *"IR Spectroscopic Imaging: from Cells to Tissue" in “Spectrochemical Analysis using Infrared Multichannel Detectors”, R.Bhargava and I.W.Levin, Editors,*. Blackwell Publishing, Sheffield, UK: 2005.

62.Steller, W.; Einenkel, J.; Horn, L. C.; Braumann, U. D.; Binder, H.; Salzer, R.; Krafft, C., Delimitation of squamous cell cervical carcinoma using infrared microspectroscopic imaging. *Anal Bioanal Chem* **2006,** 384, (1), 145-54.

63. http://seer.cancer.gov/csr/1975_2004/results_single/sect_01_table.01.pdf. **2007**.

64. Bhargava, R., *Anal Bioanal Chem (in press)*.

65.Paluszkiewicz, C.; Kwiatek, W. M., Analysis of human cancer prostate tissues using FTIR microspectroscopy and SRIXE techniques. *Journal of Molecular Structure* **2001,** 565, 329-334.

66.Bhargava, R.; Fernandez, D. C.; Schaeberle, M. D.; Levin, I. W., Theory and application of gain ranging to Fourier transform infrared spectroscopic Imaging. *Applied Spectroscopy* **2001,** 55, (12), 1580-1589.

67.Hsu, H. S.; Lin, S. Y.; Li, M. J.; Liang, R. C., Ultrastructural and biophysical studies on protein conformations of epithelium and stroma in benign prostatic hyperplasia before and after transurethral resection of the prostate. *Ultrastructural Pathology* **2002,** 26, (3), 137-141.

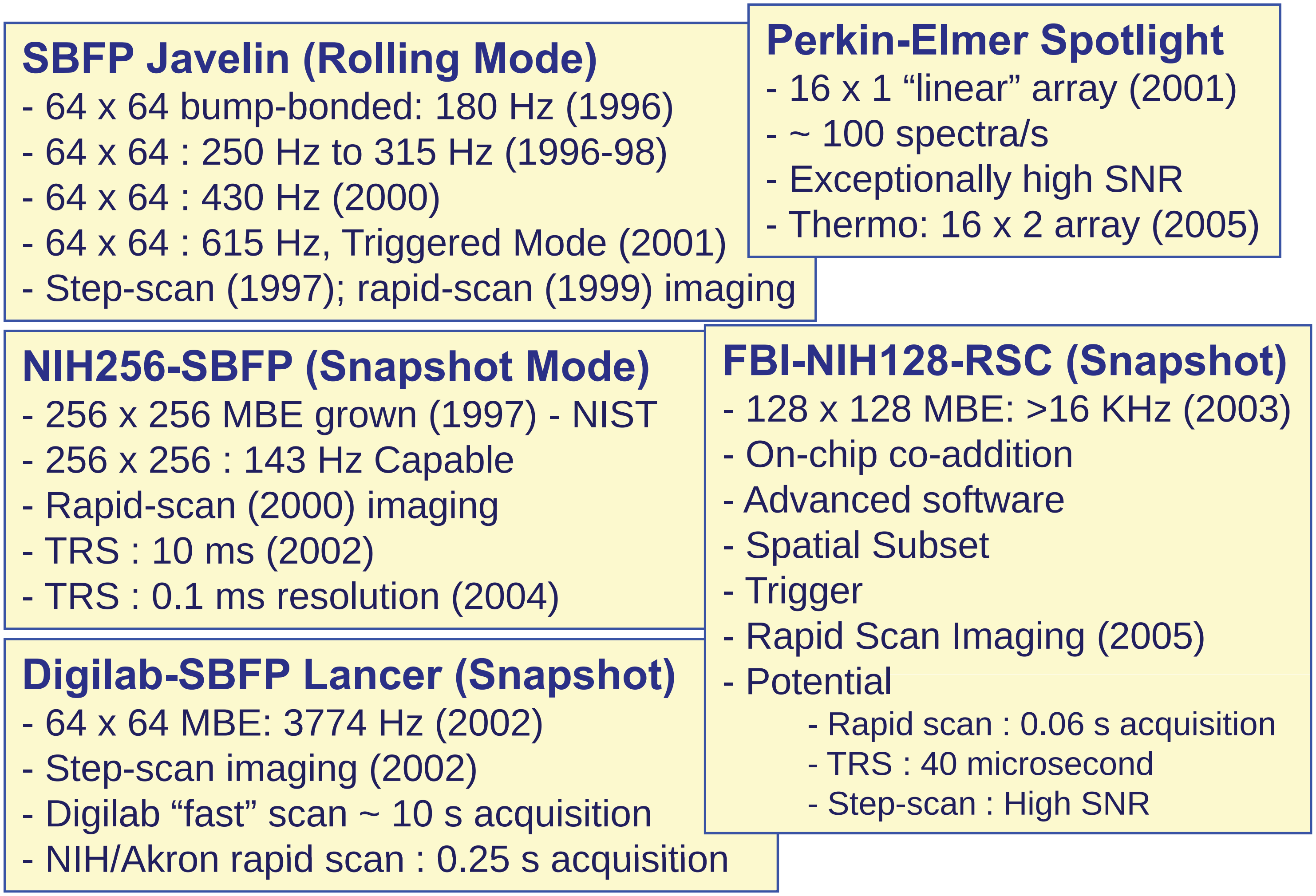
68.Li, M. J.; Hsu, H. S.; Liang, R. C.; Lin, S. Y., Infrared microspectroscopic detection of epithelial and stromal growth in the human benign prostatic hyperplasia. *Ultrastructural Pathology* **2002,** 26, (6), 365-370.

69.Gazi, E.; Dwyer, J.; Gardner, P.; Ghanbari-Siahkali, A.; Wade, A. P.; Miyan, J.; Lockyer, N. P.; Vickerman, J. C.; Clarke, N. W.; Shanks, J. H.; Scott, L. J.; Hart, C. A.; Brown, M., Applications of Fourier transform infrared microspectroscopy in studies of benign prostate and prostate cancer. A pilot study. *J Pathol* **2003,** 201, (1), 99-108.

70.Gazi, E.; Dwyer, J.; Lockyer, N.; Gardner, P.; Vickerman, J. C.; Miyan, J.; Hart, C. A.; Brown, M.; Shanks, J. H.; Clarke, N., The combined application of FTIR microspectroscopy and ToF-SIMS imaging in the study of prostate cancer. *Faraday Discuss* **2004,** 126, 41-59; discussion 77-92.

71.Gazi, E.; Dwyer, J.; Lockyer, N. P.; Miyan, J.; Gardner, P.; Hart, C.; Brown, M.; Clarke, N. W., Fixation protocols for subcellular imaging by synchrotron-based Fourier transform infrared microspectroscopy. *Biopolymers* **2005,** 77, (1), 18-30.

72.Gazi, E.; Dwyer, J.; Lockyer, N.; Gardner, P.; Miyan, J.; Hart, C.; Brown, M.; Clarke, N., A study of cytokinetic and motile prostate cancer cells using synchrotron based FTIR-microspectroscopic imaging. *Vibrational spectroscopy* **2005,** 38, (1-2), 193-201.
73.German, M. J.; Hammiche, A.; Ragavan, N.; Tobin, M. J.; Cooper, L. J.; Matanhelia, S. S.; Hindley, A. C.; Nicholson, C. M.; Fullwood, N. J.; Pollock, H. M.; Martin, F. L., Infrared spectroscopy with multivariate analysis potentially facilitates the segregation of different types of prostate cell. *Biophysical Journal* **2006,** 90, (10), 3783-3795.
74.Gazi, E.; Baker, M.; Dwyer, J.; Lockyer, N. P.; Gardner, P.; Shanks, J. H.; Reeve, R. S.; Hart, C. A.; Clarke, N. W.; Brown, M. D., A correlation of FTIR spectra derived from prostate cancer biopsies with Gleason grade and tumour stage. *European Urology* **2006,** 50, (4), 750-761.
75.Wolkers, W. F.; Balasubramanian, S. K.; Ongstad, E. L.; Zec, H. C.; Bischof, J. C., Effects of freezing on membranes and proteins in LNCaP prostate tumor cells. *Biochimica Et Biophysica Acta-Biomembranes* **2007,** 1768, (3), 728-736.
76.Bhargava, R.; Fernandez, D. C.; Hewitt, S. M.; Levin, I. W., High throughput assessment of cells and tissues: Bayesian classification of spectral metrics from infrared vibrational spectroscopic imaging data. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2006,** 1758, (7), 830-845.
77.Lasch, P.; Diem, M.; Naumann, D. In *FT-IR microspectroscopic imaging of prostate tissue sections*, Biomedical Vibrational Spectroscopy and Biohazard Detection Technologies, San Jose, CA, USA, 2004; SPIE: San Jose, CA, USA, 2004; pp 1-9.
78.Bhargava, R.; Hewitt, S. M.; Levin, I. W., Reply to Unrealistic expectations for IR microspectroscopic imaging. *Nat Biotech* **2007,** 25, (1), 31-33.



SBFP Javelin (Rolling Mode)

- 64 x 64 bump-bonded: 180 Hz (1996)
- 64 x 64 : 250 Hz to 315 Hz (1996-98)
- 64 x 64 : 430 Hz (2000)
- 64 x 64 : 615 Hz, Triggered Mode (2001)
- Step-scan (1997); rapid-scan (1999) imaging

Perkin-Elmer Spotlight

- 16 x 1 “linear” array (2001)
- ~ 100 spectra/s
- Exceptionally high SNR
- Thermo: 16 x 2 array (2005)

NIH256-SBFP (Snapshot Mode)

- 256 x 256 MBE grown (1997) - NIST
- 256 x 256 : 143 Hz Capable
- Rapid-scan (2000) imaging
- TRS : 10 ms (2002)
- TRS : 0.1 ms resolution (2004)

FBI-NIH128-RSC (Snapshot)

- 128 x 128 MBE: >16 KHz (2003)
- On-chip co-addition
- Advanced software
- Spatial Subset
- Trigger
- Rapid Scan Imaging (2005)
- Potential
  - Rapid scan : 0.06 s acquisition
  - TRS : 40 microsecond
  - Step-scan : High SNR

Digilab-SBFP Lancer (Snapshot)

- 64 x 64 MBE: 3774 Hz (2002)
- Step-scan imaging (2002)
- Digilab “fast” scan ~ 10 s acquisition
- NIH/Akron rapid scan : 0.25 s acquisition

Figure 1. Various MCT FPA detectors employed for FTIR imaging since the first reports using Santa Barbara Focalplane (SBFP) array detectors. The years in parentheses are the first reports of use for FTIR imaging. Perkin-Elmer introdcued the concept of utilizing a small linear array for very high signal to noise ratios, an approach that has since been adopted by Thermo. Our research efforts have involved the use of a high end, custom-built detector that allows for fast imaging.

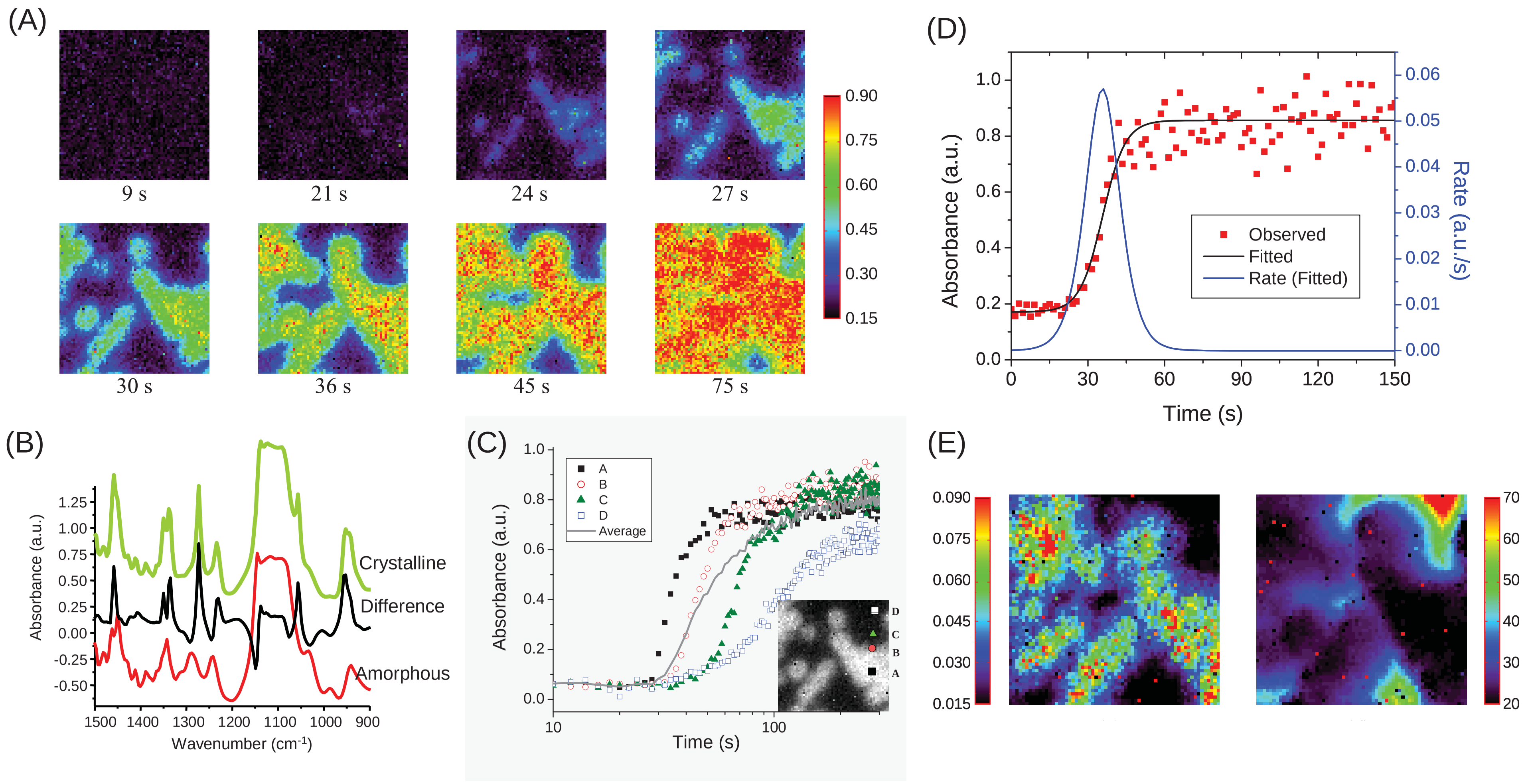


Figure 2. FTIR spectroscopy and imaging permits examination of molecular conformation changes preceding and during polymer crystallization. (A) The distribution of crystalline and amorphous fractions as a function of time for undercooling PEO ~ 13^{o} C below its melting point can be observed by the intensity of any peak that is different (B). The pixels crystallizing first can be analyzed prior to crystal formation for pre-ordering transitions. (C) Different regions of the sample have different kinetics (symbols), which are not apparent in the average spectral change (line). (D) The kinetic data (noisy)can be fit with a smooth curve and the rate of crystallization obtained. (E) spatial variation of crystallization rate (E, left) correlates with the onset of crystallization (E, right). Those regions that start to crystallize late also have a lower rate and lower ultimate purity, likely due to diffusion of impurities.

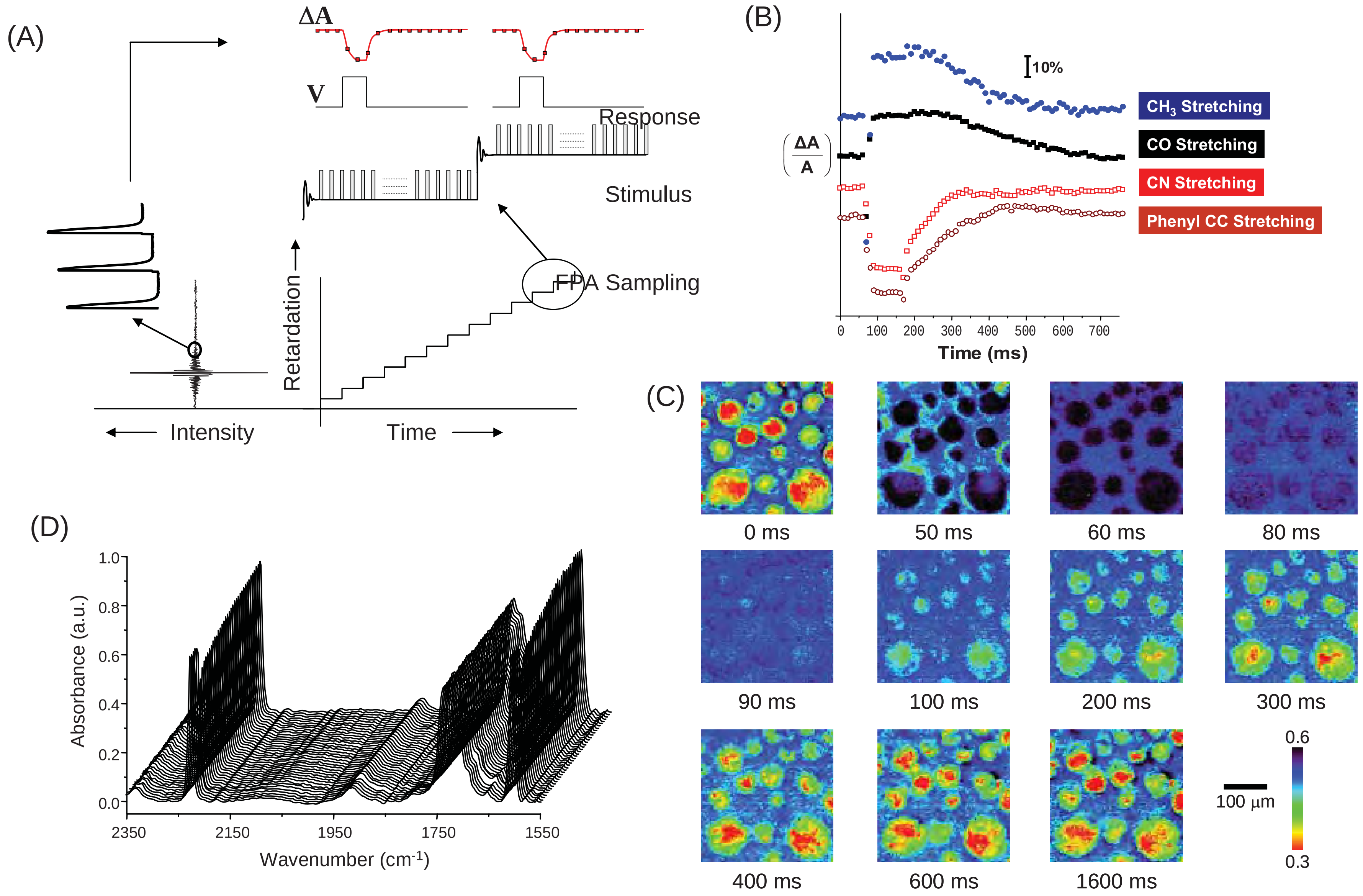


Figure 3. Time-resolved FTIR imaging can provide spatially-resolved, millisecond level dynamics over large sample areas. The operation (A) of the interferometer is similar to that of conventional step scan spectroscopy, except that an entire image is acquired for every sampling point. (B) Various functional groups may be monitored in time at specific pixels or (C) the entire image may be visualized. (D) Entire spectra from pixels may also be observed in the manner of conventional time resolved spectroscopy.

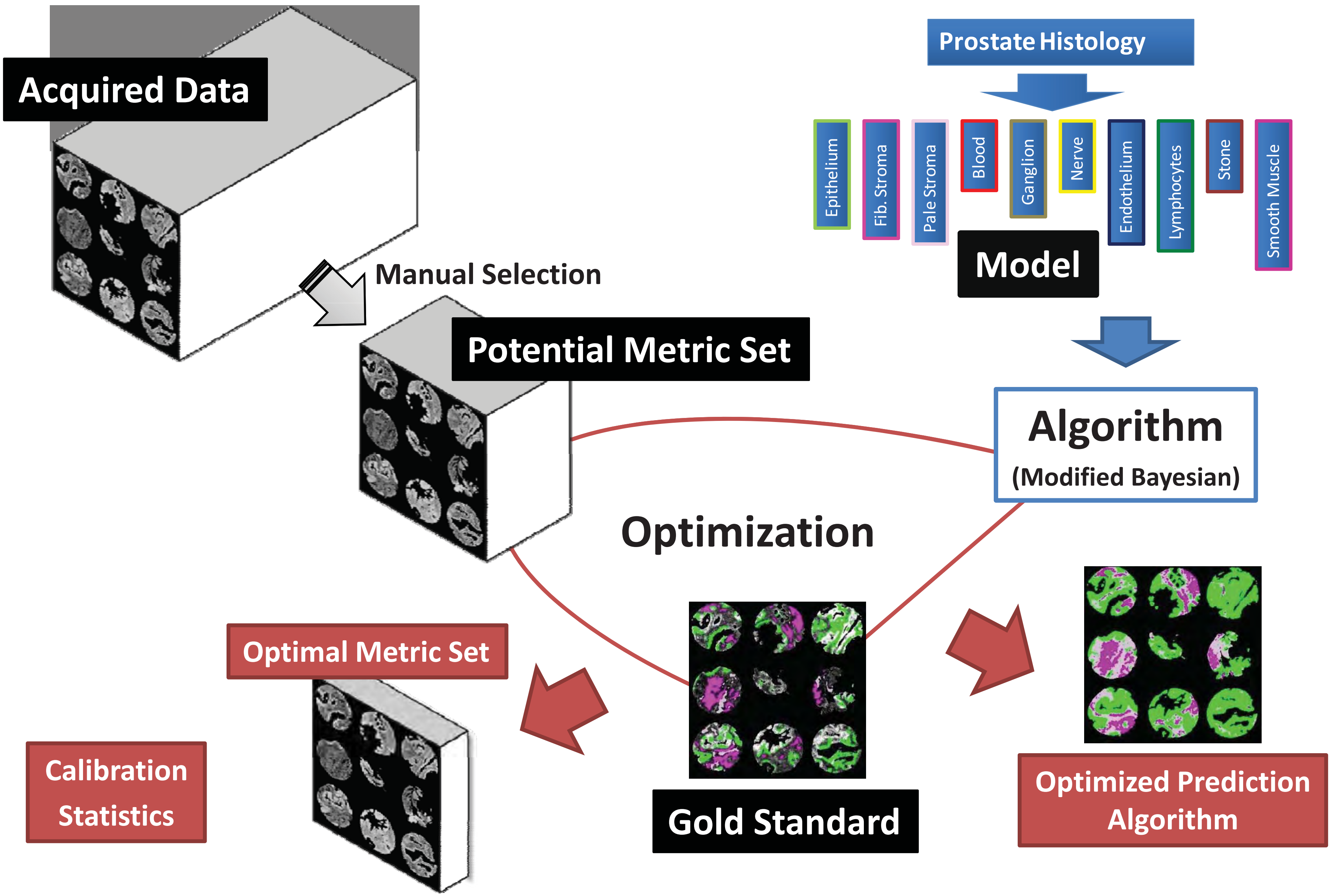
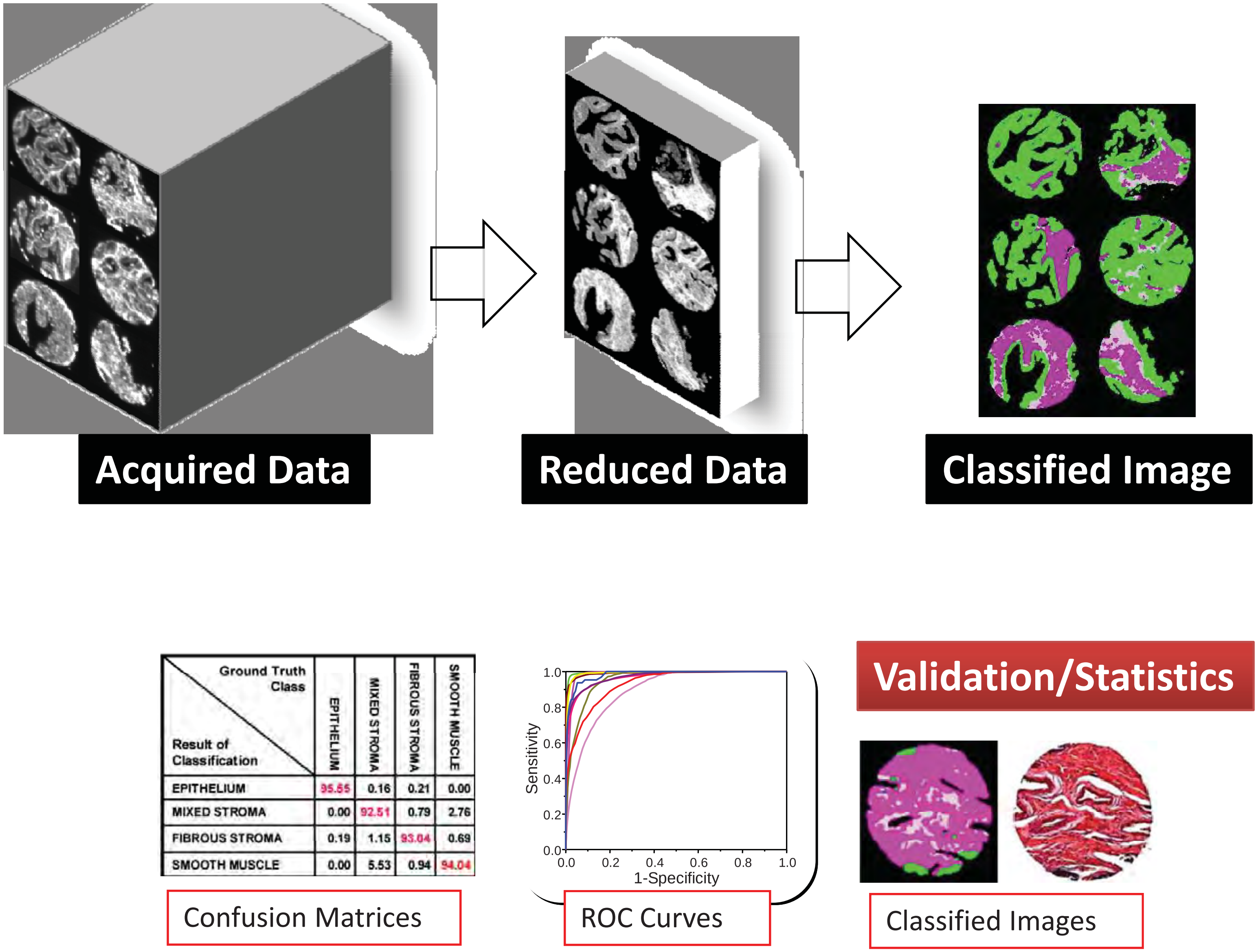


Figure 4. Organization of data into a prediction algorithm involves several steps. Acquired FTIR imaging data (top, left) is reduced by manual selection to a set of features that capture the essential elements of spectra from all tissue types. A model (top, right) is selected for the data and employed to develop an algorithm. The algorithm is applied to the entire metric set and prediction capabilities are optimized. Results of the optimization provide an optimal metric set for validation studies, the parameters of the algorithm to be applied and calibration classification statistics. The optimized algorithm is applied to acquired data without supervision (figure 5).



Result of Classification \ Ground Truth Class	EPITHELIUM	MIXED STROMA	FIBROUS STROMA	SMOOTH MUSCLE
EPITHELIUM	95.65	0.16	0.21	0.00
MIXED STROMA	0.00	92.51	0.79	2.76
FIBROUS STROMA	0.19	1.15	93.04	0.69
SMOOTH MUSCLE	0.00	5.53	0.94	94.04

Figure 5. Validation or unsupervised application of the developed protocol. FTIR imaging data are acquired (top, left), reduced to the optimal metric set (obtained as in figure 4), which is then converted to a single image that denotes each cell type by a specific color and empty space by black (top, right). Classified images can be compared to ground truth images by using confusion matrices, ROC curves and by comparisons of pixels between images. Statistical measures from these validation tests provide quantifiable results and high confidence in the development of robust algorithms.

Practical protocols for fast histopathology by Fourier transform infrared spectroscopic imaging

Frances N. Keith, Rohith K. Reddy, and Rohit Bhargava[*]
Department of Bioengineering and Beckman Institute for Advanced Science and Technology,
University of Illinois at Urbana-Champaign
3216 Digital Computer Lab, MC-278, 1304 W. Springfield Ave., Urbana, IL 61801

ABSTRACT

Fourier transform infrared (FT-IR) spectroscopic imaging is an emerging technique that combines the molecular selectivity of spectroscopy with the spatial specificity of optical microscopy. We demonstrate a new concept in obtaining high fidelity data using commercial array detectors coupled to a microscope and Michelson interferometer. Next, we apply the developed technique to rapidly provide automated histopathologic information for breast cancer. Traditionally, disease diagnoses are based on optical examinations of stained tissue and involve a skilled recognition of morphological patterns of specific cell types (histopathology). Consequently, histopathologic determinations are a time consuming, subjective process with innate intra- and inter-operator variability. Utilizing endogenous molecular contrast inherent in vibrational spectra, specially designed tissue microarrays and pattern recognition of specific biochemical features, we report an integrated algorithm for automated classifications. The developed protocol is objective, statistically significant and, being compatible with current tissue processing procedures, holds potential for routine clinical diagnoses. We first demonstrate that the classification of tissue type (histology) can be accomplished in a manner that is robust and rigorous. Since data quality and classifier performance are linked, we quantify the relationship through our analysis model. Last, we demonstrate the application of the minimum noise fraction (MNF) transform to improve tissue segmentation.

Keywords: Breast Cancer, FT-IR Spectroscopy, Hyperspectral, Histopathology, Imaging, Diagnostics, MNF Transform

1. INTRODUCTION

As histologic analysis of biopsied tissue forms the standard in definitive diagnosis of breast lesions, it is estimated that more than 1.6 million women undergo breast biopsies each year in the US alone. Biopsy samples are fixed to ensure tissue stability[1] and then sectioned for staining.[2] Microscopic examinations of stained tissue sections by a trained pathologist are the gold standard used in diagnosing breast cancer.[3] Unfortunately, these evaluations are time consuming[4] and do not always lead to an unequivocal diagnosis. For example, a study of 481 breast cancer patients from 1982-2000 at a regional cancer center indicated that 73% of ductal carcinoma *in situ* (DCIS) patients are referred by a general pathologist to an expert pathologist for review.[5] After review, 43% of these cases received different treatment recommendations. Another study found that 52% of cases referred to a multidisciplinary tumor review board received different surgery recommendations.[6] Clearly, the diagnostic process is sub-optimal. Rapid, objective second opinions are desirable. The use of emerging biological understanding and technologies for diagnoses could provide additional information in tumor evaluation and help make accurate therapy decisions. Further, it is likely that the morphologic parameters of current diagnoses are insufficient and additional information must be added. This information is typically biochemical in nature. For example, staining for human epidermal growth factor receptor 2 (HER2) can identify 25-30% of breast cancers.[7] Such examples of success, unfortunately, are uncommon for cancers in complex tissues. Hence, alternative methods are urgently required to aid diagnostic pathology.

One such means is the use of molecular spectroscopy. For example, Fourier transform infrared (FT-IR) spectroscopy is traditionally used for molecular identifications and biomolecular structure elucidations, but is not currently applied in clinical pathology.[8] An IR spectrum provides a unique molecular fingerprint with a quantitative measure of the molecular bonds present in an examined material.[9] Thus it should give a reproducible measurement of tissue

[*] rxb@uiuc.edu; phone 1 217 265 6596; cisl.bioen.uiuc.edu

Biomedical Optical Spectroscopy, edited by Anita Mahadevan-Jansen,
Wolfgang Petrich, Robert R. Alfano, Alvin Katz, Proc. of SPIE Vol. 6853,
685306, (2008) · 1605-7422/08/$18 · doi: 10.1117/12.762468

composition. Tissue, however, is microscopically heterogeneous and the measurement of chemical composition must be made in the context of knowledge of tissue structure (histology).[10] The recent emergence of FT-IR imaging couples spectroscopy and microscopy to permit rapid acquisition of spectra from tens of thousands of pixels at a high spatial resolution. Each pixel (spectrum) typically contains thousands of data points in the mid-IR wavelength region (2-12μm).[11] Automated classification can then be employed for rapid computerized tissue image analysis, as has been practiced in both the spectral processing and image processing communities. The end goal of the measurement and associated data processing steps is to permit the rapid segmentation of different types of tissue without the need for chemical dyes or contrast agents.[10] Last, the use of FT-IR imaging only involves light interacting with a sample and, unlike conventional biochemical analysis methods, does not alter the tissue in any manner. Thus it can provide additional information for pathology without the necessity of additional materials, tissue samples or changes in clinical protocols.

In this manuscript we use breast tissue as an example to illustrate the application of FT-IR imaging coupled with computerized classification for histopathology. Specifically, we demonstrate that a combination of FT-IR imaging, classification algorithms and integrated computational methods for enhancement of acquired data can be used in tandem to optimize the development of practical protocols for automated histopathology. Previous studies report on the potential for IR spectroscopy in breast pathology,[12,13,14,15,16,17] but no complete study on the spectral features of different histologic types of breast tissue exists. Preliminary efforts indicate significant spectral variation between different types of breast tissue and breast tumors,[18,19,20] but a protocol for clinical translation is lacking. We combine fast FT-IR imaging and tissue microarray sampling to demonstrate the effectiveness of our approach for automated breast histopathology on normal and malignant tissue from five patients. This approach is distinct from that in Raman spectroscopy, where histologic models are used in analyzing spectra.[21,22] As a first step towards automated tissue segmentation, we distinguish breast stroma and epithelium. This is a critical step, as over 99% of breast tumors arise in the epithelial tissue lining milk ducts and lobules.[23] False color classified images denoting stroma and epithelium are produced, followed by analysis of data collection parameters. We evaluate the impact of spectral resolution and noise on classification accuracy to demonstrate potential for faster data acquisition without loss in classification confidence. This study presents an initial effort in developing applications for FT-IR imaging in clinical pathology.

2. METHODOLOGY

2.1 Data Acquisition

The first studies to examine IR spectra of tissue began over fifty years ago,[24] but the field did not truly make progress due to limitations in instrumentation. Today, a combination of an IR microscope, Michelson interferometer and focal plane array (FPA) detector[25] permits efficient data acquisition for large sample areas. The data presented in this study is collected using the Perkin-Elmer Spotlight 400 imaging spectrometer. A spatial pixel size of 6.25 μm and a spectral resolution of 4 cm^{-1} were employed, with 2 scans averaged for each pixel. An IR background is collected with 120 scans co-added at a location on the substrate where no tissue is present. No undersampling was employing in data acquisition and a NB medium apodization function was used. A ratio of the background to tissue spectra is then computed to remove substrate and air contributions to the spectral data. The Spotlight software atmospheric correction algorithm is applied to eliminate remaining atmospheric contributions to the tissue spectra. As opposed to other configurations that employ a large FPA detector, this instrument employs a linear array detector that is raster scanned to acquire data from large sample areas. We use a combination of instrument control and post-processing software to computationally re-organize data acquired into large image sizes. Images of stained tissue are acquired using a standard Zeiss optical microscope.

2.2 Tissue sampling

Tissue microarrays (TMAs) permit facile comparison of small tissue samples from numerous patients[26] and are an especially useful sampling medium for spectroscopic analyses.[27] A TMA contains numerous small round tissue samples, termed cores, which are extracted from biopsy samples from different patients. Two paraffin-embedded TMAs were obtained from a commercial source (US Biomax) for this study. The first TMA section is placed on a glass slide and stained with hematoxylin and eosin (H&E) dyes. In H&E staining, hematoxylin stains nucleic acids and eosin stains protein-rich tissue regions. This section is used for visual morphology interpretation by a pathologist. The second TMA section is placed on a barium fluoride (BaF_2) substrate for FT-IR imaging. Though the arrays contained a large number of samples, a smaller subset of malignant and normal tissue cores from five patients with invasive ductal carcinoma (IDC) is selected for this study as the illustrative example. Each of the ten cores is 1.5 mm in diameter; hence, at a 6.25 µm pixel size, approximately 280,000 spectra are collected for each core. This results in the collection of over 560,000 spectra for each patient and approximately 2.8 million total spectra for all ten cores. This large spectral dataset facilitates rigorous validation of classification protocols at a pixel level. Paraffin is removed from the TMA by immersion in hexane with continuous stirring at 40 ^{0}C for 48-72 hours. Spectra are recorded at several locations on the TMA every 24 hours during this period to monitor paraffin removal with the disappearance of the 1462 cm^{-1} peak.

2.3 Image analysis and classification

A supervised segmentation method is used for FT-IR image classification. This algorithm has been described in detail elsewhere,[28] but is based on a modified version of a Bayesian classifier. First, the spectral profile of 1641 bands is reduced to a set of 89 useful metrics by examination of spectra from manually selected stroma and epithelium tissue regions. Metrics are manually selected to include peak ratios, peak areas, and peak centers of gravity. A metric profile M is generated for each pixel in each tissue image of the form

$$M = [m_1, m_2, m_3, \ldots m_{n_m}],\ n_m{=}89 \tag{1}$$

where each m_i is the value for a single metric and n_m is the total number of manually selected metrics. Frequency distributions for stroma and epithelium are determined for each metric and used to estimate the probability of a given metric profile representing either of these two classes. The probability of an image pixel from each class c_i being represented by a given metric profile is determined using Bayes' Rule

$$p(c_i|M) = \frac{p(M|c_i)\,p(c_i)}{p(M)} \tag{2}$$

where $p(M|c_i)$ is estimated from the metric class frequency distributions and $p(M)$ is the probability of a given metric profile. The prior probability of particular tissue class $p(c_i)$ in this model cannot be determined due the manual selection of tissue classes on FT-IR images, and is estimated as 0.5. Other ways to estimate or optimize the class prior probability may be utilized; we have noticed anecdotally, however, that the choice of this value across a large range does not significantly affect the classification results. Classification accuracy is estimated with receiver operating characteristic (ROC) analysis for selected tissue regions. The area under the ROC curve (AUC) is used to evaluate classifier sensitivity and specificity and estimate the potential of the algorithm for accurate histology determinations. The classification algorithm is trained on a large array dataset and separately validated on a second array. It is notable that we do not develop the entire classification algorithm anew here. First, the central idea of this manuscript is to demonstrate the optimization of a developed protocol and second, the sample sizes chosen here are insufficient for de novo algorithm development. Data is analyzed using the Environment for Visualizing Images (ENVI) software and with programs written in-house using Interactive Data Language (IDL).

2.4 Spectral resolution and noise analysis

Spectral resolution and noise are two common experimental variables that affect results in IR spectral analyses. The effects of spectral resolution and spectral noise are evaluated here in the context of quantitative histologic segmentation to minimize data collection time. As per the trading rules of IR spectroscopy, data collection time is expected to decrease linearly with spectral resolution and a quadratic rate with reduction in signal-to-noise ratio (SNR).[29] Ideally, these parameters would be analyzed by acquiring data at different spectral resolutions and numbers of spectral co-adds. However, the time required to collect multiple images for the TMA is prohibitive. Instead, computational methods are used to examine these parameters using the original FT-IR images acquired at 4 cm^{-1} and 2 scans per pixel. First, spectral resolution is evaluated by downsampling the data using a neighbor binning procedure to resolutions of 8, 16, 32, 64 and 128 cm^{-1}. Classification is then performed on downsampled datasets to determine the coarsest spectral resolution needed for satisfactory stroma and epithelium segmentation. For a fine spectral resolution data set at 4 cm^{-1}, the effect of noise is evaluated by adding to each spectrum noise in Gaussian distributions with standard deviations of 0.001, 0.01, and 0.1 au. Classification accuracy is estimated by evaluating the AUC at each noise standard deviation. Computational noise reduction with the minimum noise fraction (MNF) transform[30] is evaluated by reducing noise in all the data sets. Classification is performed with the same algorithm on these MNF transformed images to determine the impact of this noise reduction algorithm on stroma and epithelium segmentation.

3. DATA

The classification model presented in this manuscript involves segmentation of stroma and epithelium, which are the two most prominent tissue classes in fixed breast tissue used for pathology evaluation.[31] In practice, the recognition of epithelial cells is especially critical for cancer diagnoses, as the vast majority (>99%) of breast cancers arise in this cell type.[23] Hence, the two class model is of practical significance. While seemingly simple and practical, however, the model can potentially be confounding as stroma consists of many cell types with disparate spectral characteristics. This model was employed to develop a classifier using training data from a TMA with forty patients. Final model calibration for sixty eight tissue cores yielded an AUC value of 0.99 with an eight metric classifier.[32,33] In this study we validate this classifier with one malignant and a matched normal TMA core from a subset of five patients. As seen in Figure 1A and B, absorbance images based on spectral features closely compare with images of H&E stained tissue. Hence, using conventional pathology knowledge we can select image pixels that unequivocally correspond between the two images - representing both stroma and epithelium. These pixels are selected by examining FT-IR images at 1080 cm^{-1} to highlight asymmetric PO_2 stretching vibrations in glycoprotein in epithelium,[14] 1236 cm^{-1} to highlight CH_2 wagging vibrations associated with collagen proteins,[34] 1652 cm^{-1} to highlight C=O stretching vibrations at the protein amide I mode,[34] and 3292 cm^{-1} to highlight NH bending vibrations at the protein amide A mode (shown as an example in Figure 1B).[35] We emphasize that multiple vibrational modes must be examined in tandem and pixels identified with great care and diligence as these form the gold standard for future comparisons. Over 185,000 pixels are marked in these ten tissue cores to serve as the gold standard for ROC analysis (as shown in Figure 1C). Selecting this large set of pixels is important to achieve a reasonable sample size to accurately estimate classification potential for the entire data set. Boundary pixels are not marked to avoid errors associated with mixed pixels in FT-IR images.[27] A qualitative comparison of stained and classified images indicates that stroma and epithelium segmentation is reasonable (Figure 1D), and this is confirmed with an AUC value of 0.98 after quantitative ROC analysis. Stroma and epithelium are easily identified on false color classified images without detailed examination and interpretation. This is advantageous over traditional staining methods that require the use of chemical dyes and subsequent expert pathologist examination for evaluation.

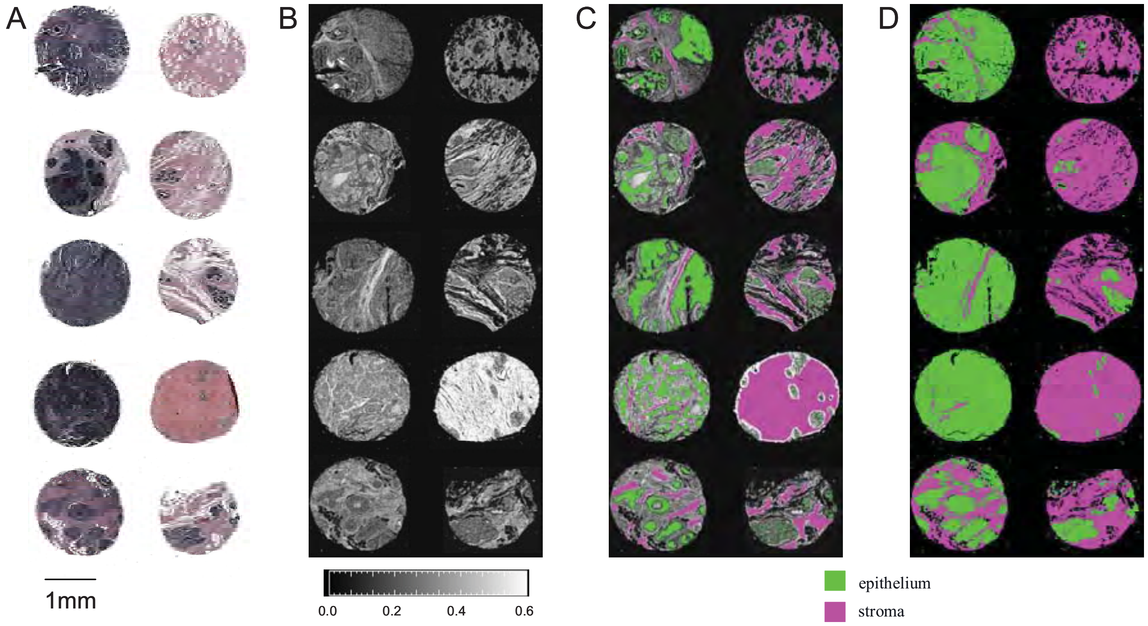


Fig. 1. Conventional H&E stained images, FT-IR spectral images and classification. (A) An H&E stained image of tissue cores from five invasive ductal carcinoma patients. Each row represents a single patient, with malignant tissue samples on the left and normal samples on the right. (B) An FT-IR image at 3292 cm^{-1} denotes the NH bending vibration at the amide A protein mode. Brighter regions denote relatively protein-rich stroma. (C) A ground truth FITR image with pixels marked as stroma or epithelium serves as the gold standard for ROC analysis and classification evaluation. (D) A classified FT-IR image in which all pixels are labeled as stroma or epithelium accurately corresponds to the H&E stained image. The classification does not require stains or human interpretation.

4. RESULTS

4.1 Effect of spectral resolution on tissue segmentation

The impact of spectral resolution on classification performance is evaluated by downsampling spectra at every pixel with a neighbor binning and interpolation procedure. FT-IR image data sets are acquired at 4 cm^{-1} spectral resolution and are downsampled to 8, 16, 32, 64, and 128 cm^{-1} resolution. As seen in Figure 2A, an average spectrum at each resolution from epithelial cells in the gold standard demonstrates that important spectral elements remain identifiable at coarser resolutions. While we anticipate that the area under the peaks would be preserved, peak shapes begin to change at a courser spectral resolution of 32 or 64 cm^{-1} due to overlaps in the complicated spectral response. It would not be surprising to note that the most robust predictors of class incorporate best both biological diversity and spectral noise (arising from both measurement and artifacts). Hence, we anticipate that the use of these metrics would also prove robust when spectra are downsampled. Figure 2B demonstrates that the classification accuracy is not significantly affected until the spectral resolution is decreased to 128 cm^{-1}.

The result is indeed surprising as numerous prior biomedical studies with vibrational spectroscopy have employed 4 cm^{-1} to 16 cm^{-1} spectral resolution. There are two important differences between the problem here and a majority of those studies. First, many of the reported studies used sensitive spectral analysis tools (e.g. second derivatives) or were looking for fine spectral features. Second, models for pathology may have needed more complex information. Here, we are examining a 2 class problem of very distinct cell types. Hence, the acceptable classification at very coarse resolutions is likely permitted by the significant biochemical differences between stroma and epithelium in the metrics selected. Previous studies have provided evidence of clear differences in IR spectra from DNA-rich tissues such as epithelium and RNA and protein-rich tissues such as stroma,[14,20] especially in the IR fingerprint region from 500-1500 cm^{-1}.[8] We hypothesize that a more complex model with additional tissue classes would likely require a higher spectral resolution for reasonable classification, but that this resolution is not required to distinguish stroma and epithelium.

A powerful feature of the algorithm we employ is the utilization of prominent spectral features for classification. Here, the features selected as classification metrics are not very sensitive to changes in spectral resolution.[36] Absorbance values are accurate if the peak full width at half maximum (FWHM) is not significantly less than the spectral resolution. As biological materials have broad and overlapping lineshapes, the condition holds even for very coarse resolutions. Therefore, the values of spectral metrics are not significantly altered even if some details in the spectrum are affected at coarser spectral resolutions. The center of gravity metrics used for classification are particularly robust, as they incorporate peak position and shape and are not strongly influenced by peak modifications in downsampled spectra. Care must be exercised in making this extrapolation to all data quality. For example, for poor signal to noise ratio spectra, the center of gravity calculation will be sensitive to noise.

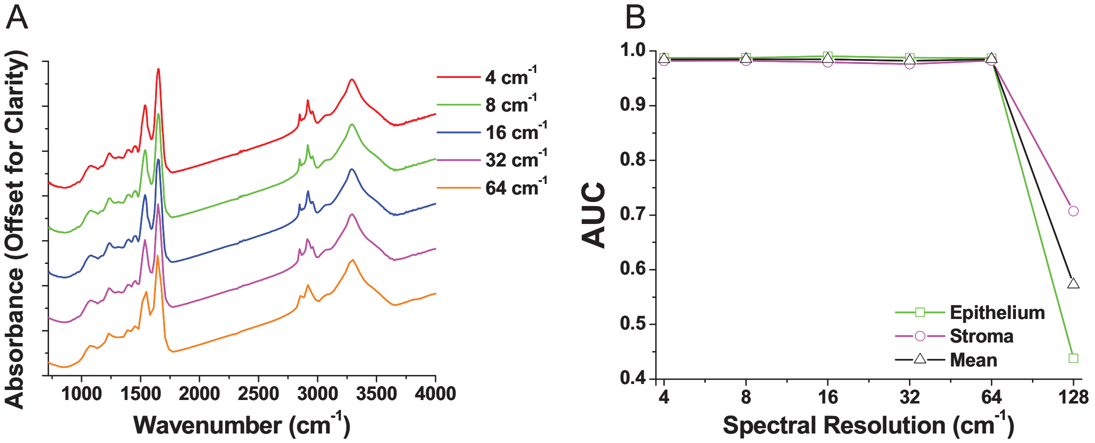


Fig. 2. Spectral resolution effect on classification. (A) Epithelial spectra obtained by downsampling data acquired at 4 cm^{-1} indicate that IR spectrum quality degrades appreciably at a spectral resolution coarser that 16 cm^{-1}, as anticipated for condensed phase biological materials. (B) AUC analysis for stroma and epithelium segmentation for each resolution demonstrates a significant decrease in classification accuracy only at a very course spectral resolution beyond 64 cm^{-1}.

The effective classification in downsampled FT-IR images presented in this manuscript indicates potential for faster data acquisition without significant loss in classification accuracy. Figure 2 suggests that no significant classification differences are observed in images up to 64 cm^{-1}. Since data acquisition time is estimated to decreased linearly with spectral resolution,[29] FT-IR images could be acquired 16 times as fast without any loss in classification performance for the two class model presented in this manuscript. Again, we emphasize that the results are preliminary and should be carefully validated. Nevertheless, the idea of optimizing data acquisition by modeling the results of other experimental conditions is an important one that should be pursued in practical translation of these protocols for clinical use.

4.2 Effect of spectral noise on tissue segmentation

Evaluation of acceptable spectral noise for FT-IR image classification is important for efficient data collection. For practical applications, it is advantageous to acquire data with the lowest SNR that permits reasonable classification. Raw data is acquired with a peak-to-peak noise value of 0.011 au, a root mean square (rms) noise value of 0.008 au, and an average amide I height of 0.328 au. To assess the impact of spectral noise on classification accuracy, Gaussian noise is added with a standard deviation of 0.001, 0.01, and 0.1 au. Figure 3 provides a qualitative evaluation of histologic images from the acquired data set (Figure 3A) and from the data sets with added Gaussian noise (Figures 3B-D).

These images indicate that acceptable classification is achieved when noise is added at a standard deviation of 0.001 au (Figure 3B), but that classification accuracy appreciably decreases with the addition of noise at or above a standard deviation of 0.01 au. This is expected, since adding noise at a standard deviation of 0.001 au does not significantly change the FT-IR image data SNR. The data set with noise added at a standard deviation of 0.01 au (Figure 3C) produces a classified image with regions of distinguishable stroma and epithelium, although there are numerous stray pixels that are not correctly classified, similar to salt and pepper noise. Upon the addition of noise of ~0.1 au, classified images become completely indistinguishable (Figure 3D), including the misidentification of many pixels on the empty region of the slides as tissue. This loss in classification accuracy is caused by an underlying broadening of spectral metric distributions for each class. This broadening bridges the difference in metric values. The overlap in values in turn decreases classification confidence as measured by the AUC. Hence, we have used the AUC as a reasonable measure of the classification accuracy at every experimental condition.

A plot of AUC against the added noise (Figure 3E) demonstrates that the AUC value remains relatively constant with the addition of low levels of noise. It then decreases to a mean AUC of 0.77 with the addition of noise at a standard deviation of 0.01 au and falls to a mean AUC of ~0.5 at a noise standard deviation of 0.1 au. It is surprising that the stroma AUC actually falls below 0.5. Though the AUC values should not be below 0.5 for classified images, our algorithm contains a pixel rejection step. A pixel is rejected if the measured metric values do not lie within the prior probability distributions. Hence, a small number of pixels are rejected at low noise levels and are not accounted.

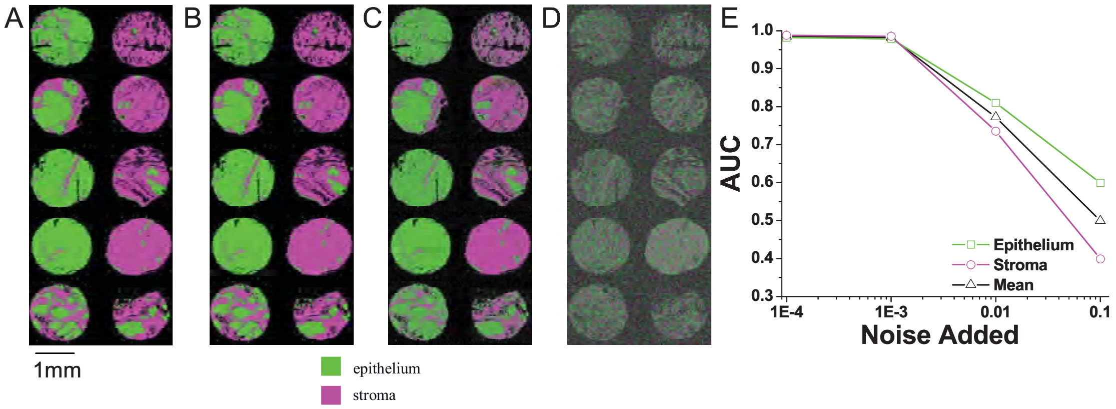


Fig. 3. Effect of noise on FT-IR image classification. Classified images are shown for (A) raw data, (B) data with Gaussian noise added at a standard deviation of 0.001 au, (C) data with Gaussian noise added at a standard deviation of 0.01 au, and (D) data with Gaussian noise added at a standard deviation of 0.1 au. (E) The AUC values for classification with noise added at a standard deviation of 0.001, 0.01, and 0.1 au confirm that classification accuracy is reasonable with a small amount of additional noise but unsatisfactory in data with a noise standard deviation at or above 0.01 au.

For the two class stroma and epithelium segmentation model presented in this manuscript, an AUC value of 0.77 does not indicate sufficient classification confidence. We would expect nearly perfect discrimination of theses two types of tissue since there are numerous spectral features that distinguish epithelium and stroma.[14,20,32,34] An estimated classification accuracy of 0.5 for this model is equivalent to random guessing and does not provide any information about tissue histology. Examination of the curve in Figure 3E indicates that some additional spectral noise at a level of 0.001 can be present without loss in classification accuracy for this two class model. We did not observe any difference in this behavior with pathology of the tissue. Breast tumor tissue is often very heterogeneous and precise pixel classification is needed to produce reasonable automated classification results. Hence these results represent a good starting point to optimize a practical protocol. There may also be a patient or clinical setting dependence of these optimal operating points that remains to be probed. From the plot, it is likely that we are close to the operating point of a practical protocol, as addition of a small amount of noise (>0.01 au) makes the classification unstable.

Last, the classification algorithm was optimized using a noise level similar to that of the acquired data set presented in this manuscript. Hence, the optimal metric sets and discriminant function are obtained for that noise level. It would not prove surprising if a de novo training and optimization of lower quality data could yield similar results. A de novo classification algorithm development, however, is not guaranteed to produce equivalent results for the higher noise cases and will fail where overlap between the prior distributions is significant due to noise broadening. Hence, we believe that the conditions found here are close to optimal.

4.3 Noise reduction with the MNF transform

In this manuscript, we have used an instrument with a high performance detector that has a low multichannel detection advantage. FT-IR imaging using large focal plane array (FPA) detectors, however, is a promising avenue for rapid data acquisitions due to the large multichannel advantage. Imaging with FPAs, unfortunately, often results in low signal-to-noise (SNR) data due to the poor detector characteristics and other limitations.[37] From the trading rules of FT-IR spectroscopy,[29] achieving a factor of n improvement in SNR would result in a increase of n^2 in data collection time. An alternative to improve SNR is to employ post-processing algorithms to reduce noise. One such avenue for noise reduction is the use of the minimum noise fraction (MNF) transform. The MNF transform can be used in a mathematical procedure to remove uncorrelated contributions from the spatial and spectral domains. First, a forward transform is used to perform a factor analysis and re-order spectral data in the order of decreasing SNR. The MNF calculation is a two-step process. A noise covariance matrix is estimated and used to decorrelate and rescale the noise in the data. Subsequently, a standard PCA performed on the noise-whitened data. A second step is to select only those factors that correspond to a sufficiently high SNR by examining the eigenvalue images. The first few eigenvalue images generally correspond to higher SNR values and contain most of the useful information. Noise reduction is achieved by suppressing the later factors corresponding largely to noise or zero-filling components and inverse transforming the data. A noise reduction by a factor greater than 5 could be achieved by this technique if the initial SNR is sufficiently high.[38,39] Though the utility of this method is demonstrated for IR imaging,[40] its use has not been widespread. Further, the use of MNF transformed data for tissue classification has not been attempted.

We propose to use the MNF transform route as a method for fast data acquisition without loss in classification accuracy. The protocol involves rapid data collection at a low SNR, followed by application of MNF transform for noise reduction. Classification is then performed on these noise-reduced images. It must be noted that the gain here is through computational techniques and does not involve changes in instrumentation hardware or data acquisition time. A secondary advantage that may arise is that decreasing the variance in spectral data could also decrease the biologic variance in the data and should improve separation of tissue classes. Excessive image noise will broaden spectral metric distributions for each class, which increases the error associated with each metric and decreases classification confidence. Therefore, if the metric distribution mean values for each class are sufficiently different decreasing noise will decrease the area of metric distribution overlap and improve segmentation confidence.

The impact of noise reduction on classification is demonstrated in Figure 4. The MNF transform-based protocol is applied to the acquired data set and the data sets with Gaussian noise added as discussed in the previous section. Classified images are displayed for each noise level after MNF transform-aided noise reduction (Figures 4A-D). The AUC values for the MNF transformed image sets are compared with the AUC values for noisy images (Figure 4E).

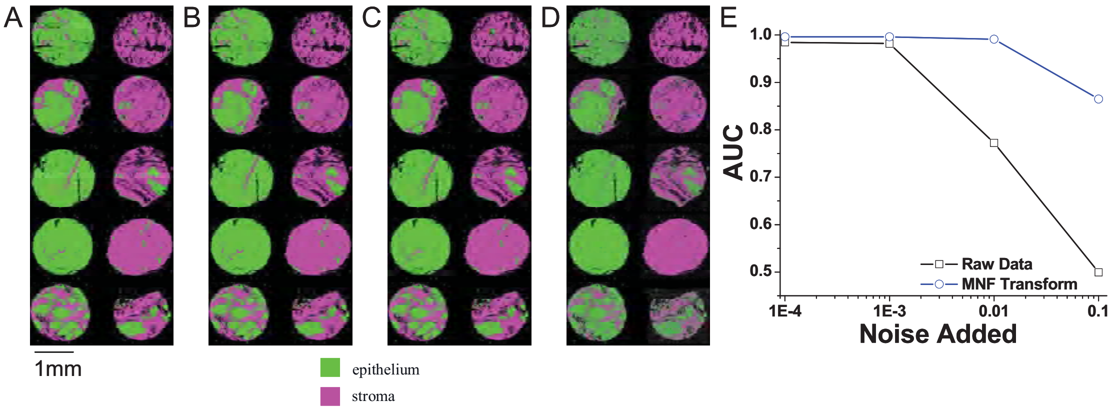


Fig. 4. Improvement in automated FT-IR image classification with the application of the MNF transform. Classified images from MNF transformed FT-IR images are shown for (A) raw data, (B) data with Gaussian noise added at a standard deviation of 0.001 au, (C) data with Gaussian noise added at a standard deviation of 0.01 au, and (D) data with Gaussian noise added at a standard deviation of 0.1 au. (E) Comparing AUC values for original FT-IR images and MNF transformed FT-IR images demonstrates that classification improves with noise reduction, especially when the noise has a standard deviation of 0.01 - 0.1 au.

Evaluation of classified images and AUC values indicates that the MNF transform improves classifier performance for each image. Given that the classification accuracy was very high, the effects of MNF transform are significant only when the noise level degrades the original data. Nevertheless, it can be seen from the figure that the high accuracy is recovered for an order of magnitude increase in data noise. Therefore, application of the MNF transform on data acquired with these noise distributions will make a significant difference in classifier performance. Specifically, we can acquire data with a noise standard deviation of 0.01 au and provide accuracy levels that are comparable to those obtained in our measurements of lower noise. This finding is significant in that noise levels of 0.01 au are commonly obtained in rapidly acquired FT-IR imaging data sets with large array detectors. Further, since the classification accuracy seems to be little affected by spectral resolution, we can anticipate that it will be little affected by the choice of an apodization function and other minor sources of error for a reasonable spectral resolution. Hence, we contend that the protocol developed here would be well-suited to rapid imaging with large array detectors.

5. CONCLUSIONS

Recent developments in FT-IR imaging and data processing facilitate new applications for this technology. In this manuscript, we report an initial application in automating histopathology of breast tissue. Supervised segmentation of breast stroma and epithelium in FT-IR images is presented and nearly-perfect classification accuracy is estimated. The impacts of spectral resolution and noise on image classification are evaluated. Results in this paper demonstrate that spectral resolution can be decreased 16-fold without loss in classification accuracy. The classification algorithm is more sensitive to noise, but noise reduction with the MNF transform can improve classification accuracy while decreasing the time required for data collection. This evaluation of the impact of experimental parameters on classification accuracy represents a first step in developing a practical protocol for rapid and automated histopathology.

REFERENCES

[1] L. Benoit, P. Fayoulet, F. Collin, L. Arnould, J. Fraisse, and J. Cuisenier, "Histological and cytopathological cancer specimens: good practice in the operating room," Ann. Chir., 128, 637-641 (2003).

[2] A. Creager and K. Geisinger, "Intraoperative evaluation of sentinel lymph nodes for breast carcinoma: current methodologies," Adv. Anat. Pathol., 9(4), 233-243 (2002).

[3] A. Cochran, R. Huang, J. Guo, and D. Wen, "Current practice and future directions in pathology and laboratory evaluation of the sentinel node," Ann. Surg. Oncol., 8(9 Suppl.), 13S-17S (2001).

[4] M. Simunovic, A. Gagliardi, D. McCready, A. Coates, M. Levine, and D. DePetrillo, “A snapshot of waiting times for cancer surgery provided by surgeons affiliated with regional cancer centers in Ontario,” CMAJ, 165(4), 421-425 (2001).

[5] E. Rakovitch, A. Mihai, J. Pignol, W. Hanna, J. Kwinter, C. Chartier, I. Ackerman, J. Kim, K. Pritchard, and L. Paszat, “Is expert breast pathology assessment necessary for the management of ductal carcinoma in situ?,” Breast Cancer Res. Treat., 87, 265-272 (2004).

[6] E. Newman, A. Guest, M. Helvie, M. Roubidoux, A. Chang, C. Kleer, K. Diehl, V. Cimmino, L. Pierce, D. Hayes, L. Newman, and M. Sabel, “Changes in surgical management resulting from case review at a multidisciplinary tumor board,” Cancer, 107, 2346-2351 (2005).

[7] D. Slamon, G. Clark, S. Wong, W. Levin, A. Ullrich, and W. McGuire, “Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene,” Science, 235(4785), 177-182 (1987).

[8] D. Naumann, “FT-infrared and FT-Raman spectroscopy in biomedical research,” Appl. Spec. Review, 36(2-3), 239-298 (2001).

[9] D. Ellis and R. Goodacre, “Metabolic fingerprinting in disease diagnosis: biomedical applications of infrared and Ramen spectroscopy,” Analyst, 131, 875-885 (2006).

[10] G. Srinivasan and R. Bhargava, “Fourier transform-infrared spectroscopic imaging: the emerging evolution from a microscopy tool to a cancer imaging modality,” Spectroscopy, 22(7), 30-43 (2007).

[11] I.W. Levin and R. Bhargava, “Fourier transform infrared vibrational spectroscopic imaging: integrating microscopy and molecular recognition,” Annu. Rev. Phys. Chem., 56, 429-474 (2005).

[12] M. Diem, M. Romeo, S. Boydston-White, M. Miljkovic, and C. Matthaus, “A decade of vibrational micro-spectroscopy of human cells and tissue (1994-2004),” Analyst, 129, 880-885 (2004).

[13] H. Fabian, N. Thi, M. Eiden, P. Lasch, J. Schmitt, and D. Naumann, “Diagnosing benign and malignant lesions in breast tissue sections by using IR-microspectroscopy,” Biochim. Biophys. Acta, 1758(7), 874-882 (2006).

[14] H. Fabian, M. Jackson, L. Murphy, P. Watson, I. Fichtner, and H.H. Mantsch, “A comparative infrared spectroscopic study of human breast tumors and breast tumor cell xenografts,” Biospectroscopy, 1, 37-45 (1995).

[15] Y. Ci, T. Gao, J. Feng and Z. Guo, “Fourier transform infrared spectroscopic characterization of human breast tissue: implications for breast cancer diagnosis,” Appl. Spec., 53(3), 312-315 (1999).

[16] C. Liu, Y. Zhang, X. Yan, X. Zhang, C. Li, W. and Yang, and D. Shi, “Infrared absorption of human breast tissues in vitro,” J. Lumin., 119-120, 132-136 (2006).

[17] C. Petibois, and G. Deleris, “Chemical mapping of tumor progression by FT-IR imaging: towards molecular histopathology,” Trends Biotechnol., 24(10), 455-462 (2006).

[18] M. Jackson, J. Mansfield, B. Dolenko, R. Somorjai, H.H. Mantsch, and P. Watson, “Classification of breast tumors by grade and steroid receptor status using pattern recognition analysis of infrared spectra,” Cancer Detect. Prev., 23(3), 245-253 (1999).

[19] R. Eckel, H. Huo, W. Guan, X. Hu, X. Che, and W.D. Huang, “Characteristic infrared spectroscopic patterns in the protein bands of human breast tissue,” Vib. Spec., 27, 165-173 (2001).

[20] M. Diem, S. Boydston-White, and L. Chiriboga, “Infrared spectroscopy of cells and tissues: shining light onto a novel subject,” Appl. Spec., 53(4), 148A-161A (1999).

[21] A. Haka, K. Shafer-Peltier, M. Fitzmaurice, J. Crowe, R.R. Dasari, and M.S. Feld, “Diagnosing breast cancer by using Raman spectroscopy,” Proc. Natl. Acad. Sci. USA, 102(35), 12371-12376 (2005).

[22] P. Matousek and N. Stone, “Prospects for the diagnosis of breast cancer by noninvasive probing of calcifications using transmission Raman spectroscopy,” J. Biomed. Opt., 12(2), 024008 (2007).

[23] C. Adem, C. Reynolds, J. Ingle, and A.. Nascimento, “Primary breast sarcoma: clinicopathologic series from the Mayo Clinic and review of the literature,” Br. J. Cancer, 91(2), 237-241 (2004).

[24] E. Blout, and R. Mellors, “Infrared Spectra of Tissues,” Science, 110, 137-138 (1949).

[25] E.N. Lewis, P.J. Treado, R.C. Reeder, G. Story, A. Dowrey, C. Marcott, and I.W. Levin, “Fourier transform spectroscopic imaging using an infrared focal-plane array detector,” Anal. Chem., 67, 3377-3381 (1995).

[26] J. Kononen, L. Bubendorf, A. Kallioniemi, M. Barlund, P. Schraml, S. Leighton, J. Torhorst, M. Mihatsch, G. Sauter, and O. Kallioniemi, “Tissue microarrays for high-throughput molecular profiling of tumor specimens,” Nat. Med., 4(7), 844-847 (1998).

[27] D.C. Fernandez, R. Bhargava, S.M. Hewitt, and I.W. Levin, “Infrared spectroscopic imaging for histopathologic recognition,” Nat. Biotechnol., 23(4), 469-474 (2005).

[28] R. Bhargava, D.C. Fernandez, S.M. Hewitt, and I.W. Levin, "High throughput assessment of cells and tissues: Bayesian classification of spectral metrics from infrared vibrational spectroscopic imaging data," Biochim. Biophys. Acta, 1758(7), 830-845 (2006).

[29] P.R. Griffiths and J. De Haseth, Fourier Transform Infrared Spectroscopy, John Wiley & Sons, New York, 1986.

[30] A. Green, M. Berman, P. Switzer, and M. Craig, "A transformation for ordering multispectral data in terms of image quality with implications for noise removal," IEEE Trans. Geosci. Remote Sensing, 26, 65-74 (1988).

[31] P.P. Rosen, Rosen's Breast Pathology, Lippincott, Williams, and Wilkins, Philadelphia, 2001.

[32]. F.N. Keith and R. Bhargava, "Data processing for tissue histopathology using Fourier transform infrared spectra," Proc. Asilomar Conference on Systems, Signals, and Computers, 71-75 (2006).

[33]. F.N. Keith and R. Bhargava, "Towards automated breast histopathology using mid-IR spectroscopic imaging," Technol. Cancer Res. Treat., in preparation.

[34] M. Jackson, L. Choo, P. Watson, W. Halliday, and H.H. Mantsch, "Beware of connective tissue proteins: assignment and implications of collagen absorptions in infrared spectra of human tissues," Biochim. Biophys. Acta, 1270, 1-6 (1995).

[35] R. Salzer, G. Steiner, H.H. Mantsch, J. Mansfield, and E.N. Lewis, "Infrared and Ramen imaging of biological and biomimetric samples," Frensenius J. Anal. Chem., 366, 712-726 (2000).

[36] R. Bhargava, "Towards a practical Fourier transform infrared chemical imaging protocol for cancer histopathology," Anal. Bioanal. Chem., 389, 1155-1169 (2007).

[37] C.M. Snively and J.L Koenig, "Characterizing the performace of a fast FT-IR imaging spectrometer," Appl. Spec., 53(2), 170-177 (1999).

[38] R. Bhargava, S.Q. Wang, and J.L. Koenig, "Route to higher fidelity FT-IR imaging." Appl. Spec., 54(4), 486-495 (2000).

[39] R. Bhargava, T. Ribar, and J.L. Koenig, "Towards faster FT-IR imaging by reducing noise," Appl. Spec., 53(11), 1313-1322 (1999).

[40] M. Wabomba, Y. Sulub, and G. Small, "Remote Detection of Volatile Organic Compounds by Passive Multispectral Infrared Imaging Measurements," Appl. Spec., 61(4), 349-358 (2007).

Anal. Chem. **2010,** *82,* 3474–3486

Theory of Midinfrared Absorption Microspectroscopy: I. Homogeneous Samples

Brynmor J. Davis,[†] P. Scott Carney,[‡] and Rohit Bhargava*[,†]

Department of Bioengineering, Department of Electrical and Computer Engineering, and the Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana–Champaign, Urbana, Illinois 61801

Midinfrared (IR) microspectroscopy is widely employed for spatially localized spectral analyses. A comprehensive theoretical model for the technique, however, has not been previously proposed. In this paper, rigorous theory is presented for IR absorption microspectroscopy by using Maxwell's equations to model beam propagation. Focusing effects, material dispersion, and the geometry of the sample are accounted to predict spectral response for homogeneous samples. Predictions are validated experimentally using Fourier transform IR (FT-IR) microspectroscopic examination of a photoresist. The results emphasize that meaningful interpretation of IR microspectroscopic data must involve an understanding of the coupled optical effects associated with the sample, substrate properties, and microscopy configuration. Simulations provide guidance for developing experimental methods and future instrument design by quantifying distortions in the recorded data. Distortions are especially severe for transflection mode and for samples mounted on certain substrates. Last, the model generalizes to rigorously consider the effects of focusing. While spectral analyses range from examining gross spectral features to assessing subtle features using advanced chemometrics, the limitations imposed by these effects in the data acquisition on the information available are less clear. The distorting effects are shown to be larger than noise levels seen in modern spectrometers. Hence, the model provides a framework to quantify spectral distortions that may limit the accuracy of information or present confounding effects in microspectroscopy.

Infrared (IR) absorption spectroscopy has been coupled to microscopy in various configurations for over 50 years.[1] The modern coupling of an interferometer with a microscope and mapping stage has enabled raster recording of Fourier transform infrared (FT-IR) spectra,[2] considerably accelerating the numbers of studies and scope of analysis by making instrumentation practical.[3] Numerous applications have been reported, for example, in materials science,[4] forensics,[5] and biomedical research.[6,7] For a number of reasons, these raster mapping systems are best utilized for point examination of specific sample areas.[8] Significantly higher imaging speeds and practical wide-field imaging can now be routinely attained by FT-IR microspectroscopy with focal plane array (FPA) detectors.[9,10] Hence, one can consider the current state of mid-IR microscopy to consist of two diverging modes. In the first, point microspectroscopy is conducted on small, homogeneous domains. The data acquisition is often guided by visible-band microscopy that is parfocal and colinear with the IR beam in the instrument; this mode is called point microscopy. The second major modality utilizes array detectors to measure across large areas of samples, with a spectrum recorded for each of tens to thousands of pixels; this mode is called imaging. The two modes are related and collectively termed microspectrosopy, in that both use focusing to collect spectra from small regions.

Ideally, FT-IR microspectroscopy can be thought to be an extension of IR spectroscopy localized to specific points in the sample. However, as FT-IR microspectroscopy is currently practiced,[3] this description is not accurate. The geometry of the sample boundaries, the morphology within the sample, the surrounding media, and the imaging optics all affect the measurements. In any given data set, the net contribution of all of these effects is observed, so that the spectra generally differ from the spectral response of the bulk material in the sample. Previous analyses of spectral differences between bulk and microscopy data have focused on the effects of stray light, the effects of oblique incidence on corrections to Beer's law[11] and orientation measurements.[4] Reports of optical distortions in FT-IR imaging have focused on the role of interfaces,[12] on scattering at an edge[13] and on scattering by the sample.[14] Distortions in a reflection–absorption (transflection)

* To whom correspondence should be addressed. E-mail: rxb@illinois.edu.

[†] Department of Bioengineering and the Beckman Institute.

[‡] Department of Electrical and Computer Engineering and the Beckman Institute.

(1) Norris, K. P. *J. Sci. Instrum.* **1954**, *31*, 284–287.
(2) Kwiatkoski, J. M.; Reffner, J. A. *Nature* **1987**, *328*, 837–838.
(3) *The Design, Sample Handling, and Applications of Infrared Microscopes*, ASTM STP 949; Roush, P. B., Ed.; American Society for Testing and Materials: West Conshohochen, PA, 1985.
(4) Koenig, J. L. *Spectroscopy of Polymers*, 2nd ed.; Elsevier: Amsterdam, The Netherlands, 1999.
(5) Bartick, E. G.; Tungol, M. W.; Reffner, J. A. *Anal. Chim. Acta* **1994**, *288*, 35–42.
(6) Wetzel, D. L.; LeVine, S. M. *Science* **1999**, *285*, 1224–1225.
(7) *Infrared and Raman Spectroscopy of Biological Materials*; Gremlich, H.-U., Yan, B., Eds.; Practical Spectroscopy; Marcel Dekker: New York, 2000.
(8) Bhargava, R.; Wall, B. G.; Koenig, J. L. *Appl. Spectrosc.* **2000**, *54*, 470–479.
(9) Levin, I. W.; Bhargava, R. *Annu. Rev. Phys. Chem.* **2005**, *56*, 429–474.
(10) Lewis, E. N.; Treado, P. J.; Reeder, R. C.; Story, G. M.; Dowrey, A. E.; Marcott, C.; Levin, I. W. *Anal. Chem.* **1995**, *67*, 3377–3381.
(11) Blout, E. R.; Bird, G. R.; Grey, D. S. *J. Opt. Soc. Am.* **1950**, *40*, 304–313.
(12) Bhargava, R.; Wang, S.-Q.; Koenig, J. L. *Appl. Spectrosc.* **1998**, *52*, 323–328.
(13) Romeo, M.; Diem, M. *Vib. Spectrosc.* **2005**, *38*, 129–132.
(14) Lee, J.; Gazi, E.; Dwyer, J.; Brown, M. D.; Clarke, N. W.; Nicholson, J. M.; Gardner, P. *Analyst* **2007**, *132*, 750–755.

10.1021/ac902067p © 2010 American Chemical Society
Published on Web 04/14/2010

measurement geometry have received particular attention,[15,14] e.g., a transformation procedure to correct a dispersive phase error has been proposed.[13] However, no study in the literature has rigorously addressed the cause of apparent spectral artifacts and morphological distortions from first principles as a function of the microscope design and sample parameters. This situation is in sharp contrast to, for example, fluorescence microscopy, where the theory is highly sophisticated and numerical corrections to the data can be confidently made.[16−20]

Considerable care must be taken in applying the methods of analysis from visible microscopy to infrared microspectroscopy. Fluorescent emissions from distinct positions within the sample are uncorrelated, for example, allowing relatively simple modeling of image formation. In the visible and near-IR spectral regions, samples typically exhibit weak and/or broad absorbance profiles and the dominant intrinsic optical process is scattering. Hence, the usual approach is to describe the sample as a collection of nondispersive scattering inhomogeneities. In the mid-IR, however, fundamental vibrational modes of molecular species are resonant with the optical frequencies of the incident radiation. These resonances lead to sharp and strong absorption features that, of course, form the very basis of spectroscopy. As a consequence, the imaginary (absorptive) part of the refractive index is significant and the real part of the index undergoes a large anomalous dispersion.[21] It is this interplay of absorption (the contrast mechanism in IR spectroscopy), anomalous dispersion, and optical energy transport that, in part, leads to complications in recording and understanding of data.

In this manuscript, rigorous optical theory is developed for IR microspectroscopy. The analysis will enable an understanding of the relationship between properties of the sample and recorded data and will enable quantitative, instrument independent, and sample-geometry independent data interpretation. While the scope is limited to point microscopy of samples with simple layered structure (i.e., no transverse variation) in this manuscript, it is demonstrated that significant spectral differences from bulk measurements and significant spectral distortions may arise. When nontrivial transverse sample structure or morphology is considered, the situation becomes more complicated and that case is addressed in the follow-up article.[22] Hence, this manuscript serves both to help in understanding the sample−instrument effects for homogeneous samples and as a basis for further development of IR microspectroscopy theory for complex sample morphologies.

The following sections first describe the development of a mathematical model for point microspectroscopy. A planewave solution of Maxwell's equations is found for the sample−instrument system, and this solution is used to construct a focused-field solution. Next, numerical simulations are presented to systematically examine the effects of focusing, dispersion, and sample structure. The model is also experimentally validated on a benchmarking sample.

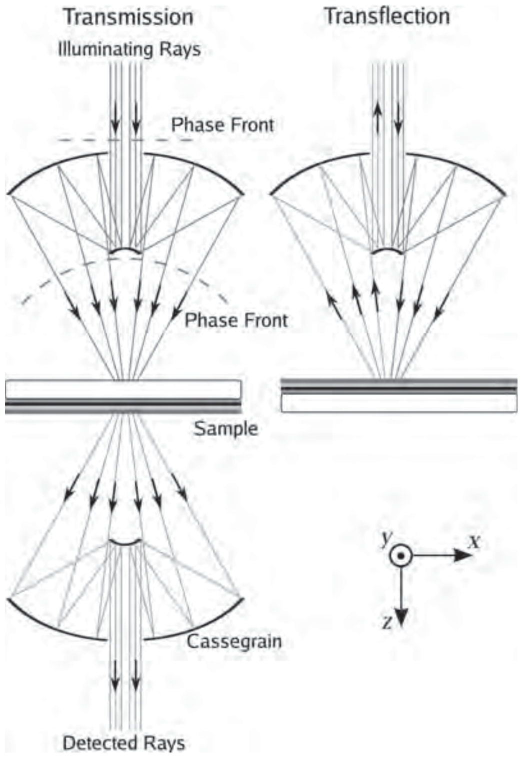


Figure 1. Illustration of focusing transmission and transflection optics. Cassegrains are used to focus light illuminating the sample and to collect the light to be detected. Loci of constant ray length are illustrated (− − −) and represent optical phase fronts. The locus above the upper Cassegrain can be regarded as an entrance pupil, and the loci below the upper Cassegrain can be regarded as an exit pupil. Note that for notational consistency, the illuminating light is always considered to originate from above the sample (i.e., from the *z* direction). The standard transmission case, where the sample is illuminated from underneath, through the substrate,[25] may be modeled using reciprocity[24] or by inverting the sample−substrate system, as illustrated. In this illustration, and in the numerical simulations that follow, the objective and condenser Cassegrains are assumed to be matched, although the theory presented is general and can account for mismatched Cassegrains.

THEORETICAL MODEL

Two generic optical systems for microspectroscopy are illustrated in Figure 1. The condensing optics focus light onto a sample supported by a planar substrate. The sample is assumed to be a layered medium without transverse structure. The resulting planar geometry, with transverse translational invariance, admits a relatively simple solution of Maxwell's equations,[23] and boundary conditions can be used to specify an incoming field consistent with focusing. As illustrated in Figure 1, transmission and transflection geometries are considered. While many IR microscopes are bottom illuminated for transmission and top illuminated for transflection, here top illumination is considered for both cases, in order to align the analytical treatment. It must be noted that the transmission case is directionally invariant

[15] Bassan, P.; Byrne, H. J.; Lee, J.; Bonnier, F.; Clarke, C.; Dumas, P.; Gazi, E.; Brown, M. D.; Clarke, N. W.; Gardner, P. *Analyst* **2009**, *134*, 1171–1175.
[16] Hiraoka, Y.; Sedat, J. W.; Agard, D. A. *Science* **1987**, *238*, 36–41.
[17] McNally, J. G.; Karpova, T.; Cooper, J.; Conchello, J. A. *Methods* **1999**, *19*, 373–385.
[18] Hell, S. W. *Nat. Biotechnol.* **2003**, *21*, 1347–1355.
[19] Gustafsson, M. G. L. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13081–13086.
[20] Rust, M. J.; Bates, M.; Zhuang, X. *Nat. Methods* **2006**, *3*, 793–795.
[21] Yamamoto, K.; Ishida, H. *Vib. Spectrosc.* **1994**, *8*, 1–36.
[22] Davis, B. J.; Carney, P. S.; Bhargava, R. *Anal. Chem.* DOI: 10.1021/ac902068e.
[23] Born, M.; Wolf, E. *Principles of Optics*, 6th ed.; Cambridge University Press: Cambridge, U.K., 1980.

for these samples,[24] hence there is no loss of generality in considering top illumination.

The electromagnetic field in each layer of the sample−instrument system may be described using planewaves which satisfy both the boundary conditions and Maxwell's equations. This type of planewave analysis[26] is commonly encountered in many fields, including the design of antireflection coatings and Fabry−Perot interferometers. In contrast to many such analyses in the visible region of the spectrum, it is necessary to consider both the real (dispersive) and imaginary (absorptive) parts of the refractive index here. The focused field is constructed as a sum of planewaves incident from the diversity of angles dictated by the focusing optics. Each planewave can be propagated through the layered sample and substrate independently in this construction, generalizing the approach of Török et al.[27] Thus, the response of the system to a single incident plane wave is addressed first, then the incident focused field is described and, finally, the total resulting measurement is predicted.

Planewave Solutions. A Cartesian coordinate system is chosen with the z axis perpendicular to the planar boundaries between sample layers. Position vectors are written $\mathbf{r} = (x, y, z)^{\mathrm{T}}$ where a superscript T denotes the vector transpose. Optical fields are represented by complex amplitudes at each temporal harmonic frequency $c\bar{\nu}$ where c is the speed of light in free space and $\bar{\nu}$ is the free space wavenumber. The permittivity and permeability of free space are denoted ε_0 and μ_0, respectively. It is assumed that the media in all layers are linear, isotropic, nonmagnetic (the relative permeability is unity), and contains no free charge. The relative permittivity $\varepsilon(\bar{\nu})$ or, equivalently, the real and imaginary (absorptive) parts of the refractive index ($n(\bar{\nu})$ and $k(\bar{\nu})$, respectively), vary from layer to layer. A single complex planewave[23] is then described by electric and magnetic fields of the respective forms

$$\mathbf{E}(\mathbf{r}, \bar{\nu}, t) = \mathbf{E}_0 \exp(i2\pi\bar{\nu}\mathbf{s}\cdot\mathbf{r}) \exp(-i2\pi\bar{\nu}ct) \tag{1}$$

$$\mathbf{H}(\mathbf{r}, \bar{\nu}, t) = \sqrt{\frac{\varepsilon_0}{\mu_0}}(\mathbf{s}\times\mathbf{E}_0) \exp(i2\pi\bar{\nu}\mathbf{s}\cdot\mathbf{r}) \exp(-i2\pi\bar{\nu}ct) \tag{2}$$

where $\mathbf{E}_0$ is the planewave amplitude vector, and $\mathbf{s}$ is a vector determining the direction of propagation and any absorptive or evanescent decay of the field. The vector $\mathbf{s} = (s_x, s_y, s_z)^{\mathrm{T}}$ obeys the dispersion relation

$$s_x^2 + s_y^2 + s_z^2 = \varepsilon(\bar{\nu}) = [n(\bar{\nu}) + ik(\bar{\nu})]^2 \tag{3}$$

For convenience, the time harmonic factors $\exp(-i2\pi\bar{\nu}ct)$ in eqs 1 and 2 are suppressed for the remainder of this article.

Samples for infrared microspectroscopy are typically mounted on a substrate and are present in air. Hence, a homogeneous sample may be considered to be a multilayer structure in which the sample, substrate, and air form a three layer system. For convenience, the effects of atmospheric absorption are neglected here. To generalize, the model system consists of L layers, each parallel to the $x-y$ plane. In each layer, the field may be written as a superposition of planewaves of the form described above, the so-called angular spectrum.[23] The electric field in the ℓth layer, that is in the z interval between the boundaries $z^{(\ell-1)}$ and $z^{(\ell)}$ (where $z^{(\ell)} > z^{(\ell-1)}$), is given by the sum over all components of the planewave angular spectrum in the slab,

$$\begin{aligned}\mathbf{E}^{(\ell)}(x, y, z, \bar{\nu}) = \bar{\nu}\int\int_{\mathbb{R}^2}\{\mathbf{B}^{(\ell)}(s_x, s_y, \bar{\nu}) \exp[i2\pi\bar{\nu}s_z^{(\ell)}(z - z^{(\ell-1)})] +\\ \hat{\mathbf{B}}^{(\ell)}(s_x, s_y, \bar{\nu}) \exp[-i2\pi\bar{\nu}s_z^{(\ell)}(z - z^{(\ell)})]\} \times\\ \exp[i2\pi\bar{\nu}(s_x x + s_y y)]\, \mathrm{d}s_x\, \mathrm{d}s_y\end{aligned} \tag{4}$$

where, by eq 3,

$$s_z^{(\ell)} = \sqrt{[n^{(\ell)}(\bar{\nu}) + ik^{(\ell)}(\bar{\nu})]^2 - s_x^2 - s_y^2} \tag{5}$$

The principal value of the square root is taken by definition, so that the downward-propagating angular spectrum $\mathbf{B}^{(\ell)}$ $(s_x, s_y, \bar{\nu})$ and the upward-propagating angular spectrum $\hat{\mathbf{B}}^{(\ell)}$ $(s_x, s_y, \bar{\nu})$ must be explicitly distinguished in eq 4. The factor of $\bar{\nu}$ is included to ensure that angular spectra constant in $\bar{\nu}$ produces a power spectrum also constant in $\bar{\nu}$. Also note that the downward propagating light, described by $\mathbf{B}^{(\ell)}$ $(s_x, s_y, \bar{\nu})$, is referenced to the upper boundary of the layer $z^{(\ell-1)}$, and the upward propagating light, described by $\hat{\mathbf{B}}^{(\ell)}$ $(s_x, s_y, \bar{\nu})$, is referenced to the lower boundary of the layer $z^{(\ell)}$.

The field in the sample is determined by the field incident from the focusing optics, i.e., by $\mathbf{B}^{(1)}(s_x, s_y, \bar{\nu})$. This field appears in eq 4 referenced to an arbitrary plane $z^{(0)}$ that does not correspond to a layer boundary but is instead chosen for convenience. Boundary conditions relate the incident field to the field in each layer of the sample and to the field in the substrate. Maxwell's equations dictate these boundary conditions and also require transversality of the field in each layer. Explicitly, Gauss' equation $\nabla\cdot\mathbf{E}(\mathbf{r}, \bar{\nu}, t) = 0$, results in the constraints

$$s_x B_x^{(\ell)}(s_x, s_y, \bar{\nu}) + s_y B_y^{(\ell)}(s_x, s_y, \bar{\nu}) + s_z^{(\ell)} B_z^{(\ell)}(s_x, s_y, \bar{\nu}) = 0 \tag{6}$$

and

$$s_x \hat{B}_x^{(\ell)}(s_x, s_y, \bar{\nu}) + s_y \hat{B}_y^{(\ell)}(s_x, s_y, \bar{\nu}) - s_z^{(\ell)} \hat{B}_z^{(\ell)}(s_x, s_y, \bar{\nu}) = 0 \tag{7}$$

The requirement that the transverse components of $\mathbf{E}(\mathbf{r}, \bar{\nu}, t)$ and $\mathbf{H}(\mathbf{r}, \bar{\nu}, t)$ are continuous across layer boundaries couples plane wave components with the same arguments $(s_x, s_y, \bar{\nu})$, across layers via the constraints

$$\begin{aligned}B_x^{(\ell)} \exp[i2\pi\bar{\nu}s_z^{(\ell)}(z^{(\ell)} - z^{(\ell-1)})] + \hat{B}_x^{(\ell)} =\\ B_x^{(\ell+1)} + \hat{B}_x^{(\ell+1)} \exp[-i2\pi\bar{\nu}s_z^{(\ell+1)}(z^{(\ell)} - z^{(\ell+1)})]\end{aligned} \tag{8}$$

[24] Potton, R. J. *Rep. Prog. Phys.* **2004**, *67*, 717–754.

[25] Carr, G. L. *Rev. Sci. Instrum.* **2001**, *72*, 1613–1619.

[26] Yeh, P. *Optical Waves in Layered Media*, 2nd ed.; Wiley-Interscience: Hoboken, NJ, 2005.

[27] Török, P.; Varga, P.; Laczik, Z.; Booker, G. R. *J. Opt. Soc. Am. A* **1995**, *12*, 325–332.

$$B_y^{(\ell)} \exp\left[i2\pi\bar{\nu}s_z^{(\ell)}\left(z^{(\ell)} - z^{(\ell-1)}\right)\right] + \hat{B}_y^{(\ell)} = B_y^{(\ell+1)} + \hat{B}_y^{(\ell+1)}\exp\left[-i2\pi\bar{\nu}s_z^{(\ell+1)}\left(z^{(\ell)} - z^{(\ell+1)}\right)\right] \tag{9}$$

$$\left(s_yB_z^{(\ell)} - s_z^{(\ell)}B_y^{(\ell)}\right) \exp\left[i2\pi\bar{\nu}s_z^{(\ell)}\left(z^{(\ell)} - z^{(\ell-1)}\right)\right] + \left(s_y\hat{B}_z^{(\ell)} + s_z^{(\ell)}\hat{B}_y^{(\ell)}\right) = \left(s_yB_z^{(\ell+1)} - s_z^{(\ell+1)}B_y^{(\ell+1)}\right) + \left(s_y\hat{B}_z^{(\ell+1)} + s_z^{(\ell+1)}\hat{B}_y^{(\ell+1)}\right) \exp\left[-i2\pi\bar{\nu}s_z^{(\ell+1)}\left(z^{(\ell)} - z^{(\ell+1)}\right)\right] \tag{10}$$

$$\left(s_z^{(\ell)}B_x^{(\ell)} - s_xB_z^{(\ell)}\right) \exp\left[i2\pi\bar{\nu}s_z^{(\ell)}\left(z^{(\ell)} - z^{(\ell-1)}\right)\right] + \left(-s_z^{(\ell)}\hat{B}_x^{(\ell)} - s_x\hat{B}_z^{(\ell)}\right) = \left(s_z^{(\ell+1)}B_x^{(\ell+1)} - s_xB_z^{(\ell+1)}\right) + \left(-s_z^{(\ell+1)}\hat{B}_x^{(\ell+1)} - s_x\hat{B}_z^{(\ell+1)}\right) \exp\left[-i2\pi\bar{\nu}s_z^{(\ell+1)}\left(z^{(\ell)} - z^{(\ell+1)}\right)\right] \tag{11}$$

For fixed arguments $(s_x, s_y, \bar{\nu})$ there are $6L$ unknowns, $3L$ for $\mathbf{B}^{(\ell)}$ $(s_x, s_y, \bar{\nu})$ and $3L$ for $\hat{\mathbf{B}}^{(\ell)}$ $(s_x, s_y, \bar{\nu})$. The transversality conditions of eqs 6 and 7 provide $2L$ linearly independent equations (one pair for each layer) and the boundary conditions of eqs 8−11 provide $4(L-1)$ linearly independent equations (four equations for each boundary). The remaining degrees of freedom allow for the specification of the incident (incoming) field at the top and bottom layers. At the last boundary, the $z = z^{(L-1)}$ plane, the field is assumed to be strictly outgoing, i.e., the incoming field is zero. Thus, it is required that

$$\hat{\mathbf{B}}^{(L)}(s_x, s_y, \bar{\nu}) = \mathbf{0} \tag{12}$$

As a result, there are only two degrees of freedom in the system, which are identified with the electric field amplitude of the illumination.

The total field is linear in the values of the illuminating field $\mathbf{B}^{(1)}(s_x, s_y, \bar{\nu})$; hence, it is instructive to consider as an example the case of single-planewave illumination, as shown in Figure 2. Notice that the coherent superposition of transmitted and reflected fields produces interference patterns in the sample and that the absorption in the sample results in decaying amplitudes into the media. These effects are also important in the case where the incident field consists of a superposition of planewaves that form a focus.

As seen in Figure 2, enforcing eqs 6−11 results in a solution where transmission, reflection, and interference effects all play a role. However, if two boundaries are separated by a large distance, the exponential factors in eqs 8−11 will result in a solution that varies rapidly with a small change in the wavelength, i.e., the interference effects will change rapidly as a function of $\bar{\nu}$. Such a situation arises when light propagates through a mounting substrate with a thickness much greater than a wavelength. This type of highly oscillatory spectral behavior will not be resolved by the spectrometer, meaning that the interference effects from the thick layer will not be observed in the data. Hence, in transmission mode, the effect of the mounting substrate can be accurately described by modeling the distant substrate−air boundary as uncoupled to the closely spaced boundaries, i.e., those associated with the sample. Thus eqs 6−11 need only to be solved for closely spaced boundaries (the air−sample and the sample−substrate boundaries), and the resulting field of interest can otherwise be propagated through the distant boundary using standard transmission coefficients.

Focused Illumination. While the previous subsection has illustrated the interaction of planewave fields with a sample, microspectroscopy involves the use of focusing optics to localize signal and increase local throughput. Focusing optics can be modeled using geometrical optics techniques such as ray tracing. In this paradigm, the Cassegrain arrangement that is usually employed for focusing in the mid-IR maps a ray on the entrance pupil to a focused ray on the exit pupil as illustrated in Figure 3. It should be noted that the locus described as the exit pupil will intersect rays emerging from the Cassegrain when the Cassegrain is used as a condenser but will intersect incoming rays when the Cassegrain is used for collection (i.e., as an objective).

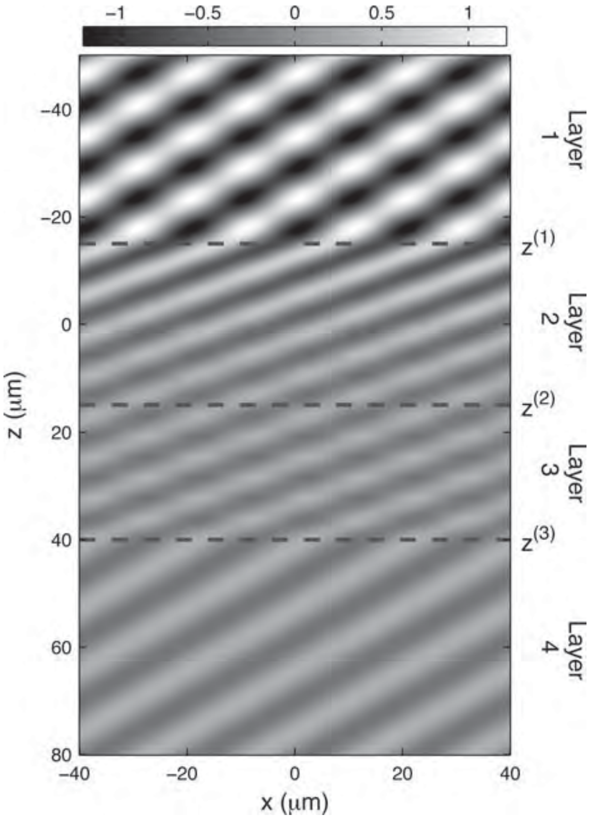


Figure 2. An example of the field produced in a layered sample under unit-amplitude planewave illumination. The illuminating light is incident at an angle of 45° in the $x-z$ plane, is purely y-polarized and has a wavelength of 10 μm in free space. The real part of the complex representation of the field is displayed. The indices of the four media present are, from top to bottom, 1, 1.4 + 0.05i, 1.4, and 1. The boundaries of the media are marked with dashed lines.

The angular spectrum amplitudes of the focused, illuminating field, $\mathbf{B}^{(1)}(s_x, s_y, \bar{\nu})$, can be associated with rays in the exit pupil.[28] As illustrated in Figure 3, the vector elements s_x and s_y determine not only the propagation direction of a focused ray but also the intersection of the associated ray path and the entrance pupil. The field in the pupil can therefore be expressed as a vector function $\mathbf{P}(s_x, s_y, \bar{\nu})$. A matrix $C_I(s_x, s_y, \bar{\nu})$ relates $\mathbf{P}(s_x, s_y, \bar{\nu})$ to $\mathbf{B}^{(1)}(s_x, s_y, \bar{\nu})$ and explicitly accounts for the optical elements (i.e., the Cassegrain) in the system,

$$\mathbf{B}^{(1)}(s_x, s_y, \bar{\nu}) = C_I(s_x, s_y, \bar{\nu})\mathbf{P}(s_x, s_y, \bar{\nu}) \tag{13}$$

[28] Wolf, E. *Proc. R. Soc. London, Ser. A: Math. Phys. Sci.* **1959**, *253*, 349–357.

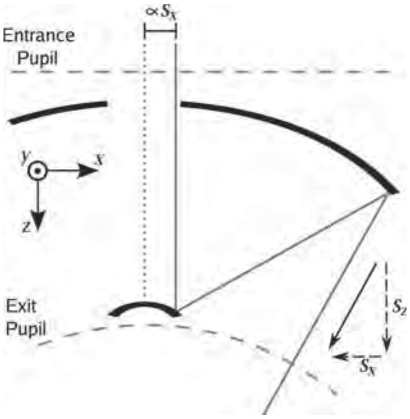


Figure 3. An illustrative ray path through a Cassegrain system. Mirrors (heavy lines) reflect rays parallel to the z axis at the entrance pupil to rays at the exit pupil which are directed to the focal point. The vector **s** gives the propagation direction for a ray and, for an aplanatic system, the transverse component of this vector (s_x in this two-dimensional figure) is proportional to the transverse position at which the ray intersects the entrance pupil. The ray path for $s_x = 0$ is represented by the dotted line; in a Cassegrain, this ray does not contribute to the focused field.

The case of a lossless aplanatic focusing system has been addressed by Richards and Wolf,[29] and results from their work can be used to find $C_I(s_x, s_y, \bar{\nu})$.

The construction of $C_I(s_x, s_y, \bar{\nu})$ is most easily accomplished by defining polarization basis vectors before and after the Cassegrain, namely, the transverse-electric (s-polarized) and transverse-magnetic (p-polarized) vectors. Assuming that the Cassegrain(s) is in free space, these vectors are

$$\mathbf{v}_s' = \mathbf{v}_s = \frac{1}{\sqrt{s_x^2 + s_y^2}}(-s_y, s_x, 0)^{\mathrm{T}} \tag{14}$$

$$\mathbf{v}_p = \frac{1}{\sqrt{s_x^2 + s_y^2}}(-s_x, -s_y, 0)^{\mathrm{T}} \tag{15}$$

$$\mathbf{v}_p' = \frac{1}{\sqrt{s_x^2 + s_y^2}}(-s_x s_z^{(1)}, -s_y s_z^{(1)}, s_x^2 + s_y^2)^{\mathrm{T}} \tag{16}$$

where a prime indicates a vector on the exit pupil side of the Cassegrain and no prime indicates the entrance pupil side. Since the focusing is performed in free space and only propagating waves are produced, $s_x^2 + s_y^2 \leq 1$, and s_x, s_y, and $s_z^{(1)}$ are all real.

The field on the exit pupil $\mathbf{P}'(s_x, s_y, \bar{\nu})$ can be found by mapping each ray through the focusing optics and correctly accounting for conservation of energy.[29] With neglect of the constant phase factors,

$$\mathbf{P}'(s_x, s_y, \bar{\nu}) = \sqrt{s_z^{(1)}}[\mathbf{v}_s'\mathbf{v}_s^{\mathrm{T}} + \mathbf{v}_p'\mathbf{v}_p^{\mathrm{T}}]\mathbf{P}(s_x, s_y, \bar{\nu}) \tag{17}$$

The field on the exit pupil can then be used to determine the resulting angular spectrum[28]

$$\mathbf{B}^{(1)}(s_x, s_y, \bar{\nu}) = \frac{\zeta\mathbf{P}'(s_x, s_y, \bar{\nu})}{s_z^{(1)}} \tag{18}$$

where ζ is the focal length of the Cassegrain. The description of the focusing optics then takes the form,

$$C_I(s_x, s_y, \bar{\nu}) = f_s(s_x, s_y, \bar{\nu})\mathbf{v}_s'\mathbf{v}_s^{\mathrm{T}} + f_p(s_x, s_y, \bar{\nu})\mathbf{v}_p'\mathbf{v}_p^{\mathrm{T}} \tag{19}$$

and in the lossless aplanatic case,

$$f_s(s_x, s_y, \bar{\nu}) = f_p(s_x, s_y, \bar{\nu}) \propto \frac{\zeta}{\sqrt{s_z^{(1)}}} \tag{20}$$

More generally, $f_s(s_x, s_y, \bar{\nu})$ and $f_p(s_x, s_y, \bar{\nu})$ can be modified to capture losses, aberrations, and the central obstruction in the Cassegrain. Note that it is implicit in this treatment that the illumination reference plane $z^{(0)}$ is the focal plane for a focus formed in free space. It can also be seen, from eq 19 that $\mathbf{B}^{(1)}(s_x, s_y, \bar{\nu})$ obeys the transversality condition of eq 6. Examples of focused angular spectra, with the central Cassegrain obstruction included, are shown in Figure 4.

In free space, the illuminating angular spectrum $\mathbf{B}^{(1)}(s_x, s_y, \bar{\nu})$ completely defines the field. The presence of the layered sample alters the field in a manner that may be calculated for each planewave component separately, as described above. The resultant focused field is then found by summing the planewave contributions in the resulting angular spectra. An example of a field focused into a layered sample is shown in Figure 5.

The analysis to this point has addressed a planewave normally incident on the entrance of the condenser Cassegrain. At close to normal incidence, a slightly off-axis illumination results in the field

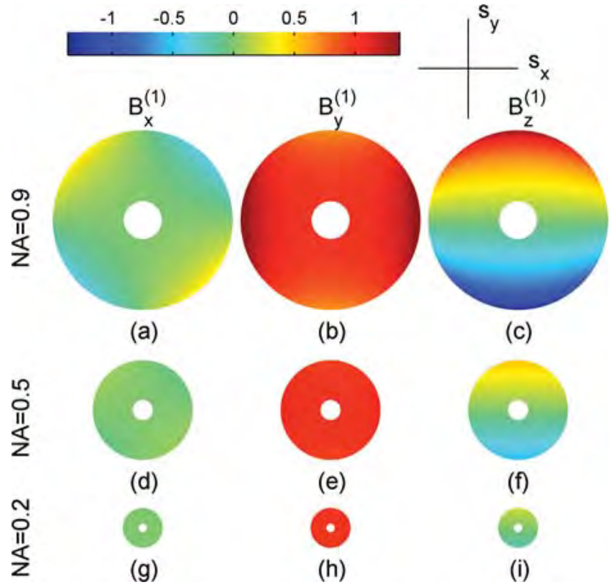


Figure 4. Normalized angular spectra $\mathbf{B}^{(1)}(s_x, s_y, \bar{\nu})$ resulting from a y-polarized planewave on the entrance pupil. The (a,d,g) x components, (b,e,h) y components, and (c,f,i) z components of $\mathbf{B}^{(1)}(s_x, s_y, \bar{\nu})$ are plotted separately. Focusing numerical apertures (NAs) of (a−c) 0.9, (d−f) 0.5, and (g−i) 0.2 are illustrated, and in each case the NA of the central obstruction is 20% of the total NA. Note that for large apertures, the y-polarized field on the entrance pupil produces significant x- and z-directed fields on the exit pupil. In transflection mode, one-half of the apertures above would be used for illumination, with the other half reserved for collection.

[29] Richards, B.; Wolf, E. *Proc. R. Soc. London, Ser. A: Math. Phys. Sci.* **1959**, *253*, 358–379.

$$\mathbf{P}(s_x, s_y, \bar{\nu}) \approx \mathbf{P}_0(s_x, s_y, \bar{\nu}) \exp[-i2\pi\bar{\nu}(s_x x_o + s_y y_o)] \tag{21}$$

at the entrance pupil, where x_o and y_o determine the inclination of the beam and the normally incident field is $\mathbf{P}_0(s_x, s_y, \bar{\nu})$. Carrying the illumination of eq 21 through eqs 13 and 4 shows that the inclination will have the effect of spatially displacing the focused field by $\mathbf{r}_o = (x_o, y_o, 0)^T$. In this manner, the angle of incidence of light on the entrance pupil governs the transverse position of the focused field. The use of aplanatic optics gives minimal distortions in the translated field.[23]

In widefield imaging, light is incident on the illumination Cassegrain at a range of angles simultaneously. Fields associated with distinct illumination angles are generally statistically uncorrelated, meaning that each can be considered individually. The resultant intensities on the detector (see the following subsection) sum, without interfering, in the process of data collection. Similarly, for unpolarized illumination, an *x*-polarized illumination field and a *y*-polarized illumination are present simultaneously. Each of these can also be analyzed independently and the measured intensity of each summed (an incoherent sum) to give the total signal.

Detection. The field at the detector may be related to the field emerging from the sample in much the same way that the illuminating field is found from the field in the entrance pupil. In transmission mode, the field exiting the Cassegrain objective, denoted $\mathbf{Q}(s_x, s_y, \bar{\nu})$, is dependent on the emerging angular spectrum $\mathbf{B}^{(L)}(s_x, s_y, \bar{\nu})$. Similar to eq 13,

$$\mathbf{Q}(s_x, s_y, \bar{\nu}) = C_T(s_x, s_y, \bar{\nu})\mathbf{B}^{(L)}(s_x, s_y, \bar{\nu}) \tag{22}$$

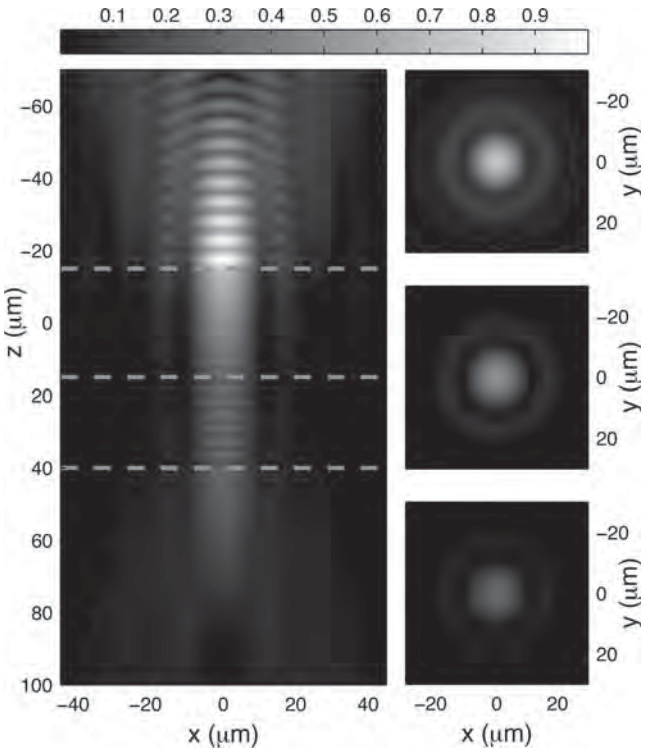


Figure 5. The focused field magnitude $|\mathbf{E}(x, y, z, \bar{\nu})|$ for $\bar{\nu} = 1000$ cm^{-1} (a free space wavelength of 10 μm), the four-layer object of Figure 2 and the angular planewave spectrum of Figure 4d–f plotted on a normalized scale. The free space focal plane is at $z^{(0)} = 0$. One transverse–axial (x–z) section is plotted at $y = 0$ and three transverse–transverse (x–y) sections are plotted at $z = -15$ μm, $z = 0$, and $z = 15$ μm. In this example, the Cassegrain pupil is filled, i.e., there are no apertures limiting the width of the illuminating beam before the focusing optics.

In the transflection mode, the field exiting the sample is the upward propagating field determined by the angular spectrum $\hat{\mathbf{B}}^{(1)}(s_x, s_y, \bar{\nu})$ and

$$\mathbf{Q}(s_x, s_y, \bar{\nu}) = C_R(s_x, s_y, \bar{\nu})\hat{\mathbf{B}}^{(1)}(s_x, s_y, \bar{\nu}) \tag{23}$$

As with the illumination matrix $C_I(s_x, s_y, \bar{\nu})$, the matrices $C_T(s_x, s_y, \bar{\nu})$ and $C_R(s_x, s_y, \bar{\nu})$ describe the focusing optics for each case. In transmission, $C_T(s_x, s_y, \bar{\nu})$ describes focusing of the downward propagating light transmitted through the sample and substrate, while in transflection $C_R(s_x, s_y, \bar{\nu})$ describes the focusing of the upward propagating reflected light.

The angular spectra emerging from the sample define the field incident on the objective Cassegrain. Similar to eq 18, this incident field can be expressed as $\mathbf{Q}'(s_x, s_y, \bar{\nu}) = \mathbf{B}^{(L)}(s_x, s_y, \bar{\nu})s_z^{(L)}/\zeta$ in transmission mode and as $\mathbf{Q}'(s_x, s_y, \bar{\nu}) = \mathbf{B}^{(1)}(s_x, s_y, \bar{\nu})s_z^{(1)}/\zeta$ in transflection mode. The mapping of the diverging field $\mathbf{Q}'(s_x, s_y, \bar{\nu})$ through an ideal objective Cassegrain to the collimated field $\mathbf{Q}(s_x, s_y, \bar{\nu})$ obeys the same relation as was given in the illumination case, i.e., eq 17. Assuming that the last layer of the system is free-space, the transmission mode relation $C_T(s_x, s_y, \bar{\nu})$ may therefore be represented compactly in the bases defined in eqs 14–16,

$$C_T(s_x, s_y, \bar{\nu}) = f_s'(s_x, s_y, \bar{\nu})\mathbf{v}_s(\mathbf{v}_s')^T + f_p'(s_x, s_y, \bar{\nu})\mathbf{v}_p(\mathbf{v}_p')^T \tag{24}$$

where

$$f_s'(s_x, s_y, \bar{\nu}) = f_p'(s_x, s_y, \bar{\nu}) \propto \frac{\sqrt{s_z^{(1)}}}{\zeta} \tag{25}$$

for the ideal Cassegrain objective. Notice that for transmission with no sample or substrate (the empty instrument case), $\mathbf{B}^{(L)}(s_x, s_y, \bar{\nu}) = \mathbf{B}^{(1)}(s_x, s_y, \bar{\nu})$, leading to the result $\mathbf{P}(s_x, s_y, \bar{\nu}) = \mathbf{Q}(s_x, s_y, \bar{\nu})$. This is to be expected; with no sample or substrate, propagation through the focusing system has no net effect.

In transflection mode a similar relation holds,

$$C_R(s_x, s_y, \bar{\nu}) = \hat{f}_s'(s_x, s_y, \bar{\nu})\hat{\mathbf{v}}_s(\hat{\mathbf{v}}_s')^T + \hat{f}_p'(s_x, s_y, \bar{\nu})\hat{\mathbf{v}}_p(\hat{\mathbf{v}}_p')^T \tag{26}$$

where $\hat{f}_s'(s_x, s_y, \bar{\nu})$ and $\hat{f}_p'(s_x, s_y, \bar{\nu})$ are as in eq 25 for ideal collection, and the reflected *s*- and *p*-polarized basis vectors are given by the expressions

$$\hat{\mathbf{v}}_s = \hat{\mathbf{v}}_s' = \frac{1}{\sqrt{s_x^2 + s_y^2}}(-s_y, s_x, 0)^T \tag{27}$$

$$\hat{\mathbf{v}}_p = \frac{1}{\sqrt{s_x^2 + s_y^2}}(s_x, s_y, 0)^T \tag{28}$$

$$\hat{\mathbf{v}}_p' = \frac{1}{\sqrt{s_x^2 + s_y^2}}(s_x s_z^{(1)}, s_y s_z^{(1)}, s_x^2 + s_y^2)^T \tag{29}$$

As before, a prime indicates a vector on the sample side of the Cassegrain.

To achieve magnification, the field $\mathbf{Q}(s_x, s_y, \bar{\nu})$ is focused on to a detector by a low-angle focusing system. For an imaging system with magnification of M, the field on the detector plane is given by

$$\mathbf{D}(x, y, \bar{\nu}) = \frac{\bar{\nu}}{M} \int \int_{\mathbb{R}^2} \mathbf{Q}(s_x, s_y, \bar{\nu}) \exp(i2\pi\bar{\nu} s_z^{(1)} z_{\mathrm{d}}) \times \exp\left[i2\pi\bar{\nu}\left(\frac{s_x}{M}x + \frac{s_y}{M}y\right)\right] \mathrm{d}s_x \, \mathrm{d}s_y, \qquad (30)$$

where $s_z^{(1)}$ is calculated as in eq 5 but with magnified values s_x/M and s_y/M instead of s_x and s_y, and z_{d} is the offset between the detector plane and the focal plane. The focusing described in eq 30 is of the same form as in eq 4 but with the focused spectrum corresponding to $\mathbf{Q}(s_x, s_y, \bar{\nu})$, the field incident on the detector focusing optics. This is justified for the focusing onto the detector, as the fields on the entrance and exit pupils of the low-angle focusing system are approximately equal, i.e., the focusing tensor $C(s_x, s_y, \bar{\nu})$ is modeled as the identity operator. Note that $\mathbf{D}(x, y, \bar{\nu})$, the field incident on the detector, lies in the $x-y$ plane as $\mathbf{Q}(s_x, s_y, \bar{\nu})$ is spanned by $\mathbf{v}_s$ and $\mathbf{v}_p$ (see eqs 14, 15, 24, and 26).

The signal measured by an optical detector is proportional to the intensity of the field integrated over the detector area, i.e.,

$$\mathbb{I}(\bar{\nu}) = \int \int_{\Omega} |\mathbf{D}(x, y, \bar{\nu})|^2 \, \mathrm{d}x \, \mathrm{d}y \qquad (31)$$

where Ω is the detector area. If the detector area is large compared to the focal spot, then the region of integration above can be replaced with the entire $x-y$ plane. In this case, Parseval's theorem can be applied to calculate the data in terms of the collimated beam exiting the Cassegrain objective, i.e., the data become independent of the focusing on to the detector with

$$\mathbb{I}(\bar{\nu}) = \int \int_{\mathbb{R}^2} |\mathbf{Q}(s_x, s_y, \bar{\nu})|^2 \, \mathrm{d}s_x \, \mathrm{d}s_y \qquad (32)$$

Here the recorded signals are simply the total intensity of the collimated beam emerging from the collection Cassegrain.

Relating Theory to Current Practice. The molecular interpretation of recorded data in microspectroscopy typically follows that of bulk spectroscopy, in which the recorded signal intensity is often interpreted using the expression

$$\mathbb{I}_S(\bar{\nu}) = |\mathbf{P}(\bar{\nu})|^2 T_S(\bar{\nu}) \exp[-4\pi\bar{\nu}k(\bar{\nu})b] \qquad (33)$$

Here $\mathbf{P}(\bar{\nu})$ is the illumination field amplitude, b is the effective path length (nominally, the sample thickness for transmission and twice the sample thickness for the transflection measurements), and $T_S(\bar{\nu})$ describes a net transmission or reflection coefficient. To calculate absorption spectra, a background measurement is first obtained,

$$\mathbb{I}_0(\bar{\nu}) = |\mathbf{P}(\bar{\nu})|^2 T_0(\bar{\nu}) \qquad (34)$$

where $T_0(\bar{\nu})$ describes the transmission and reflection effects for the experimental setup without the sample. The recorded absorbance, $A(\bar{\nu})$, is obtained from the normalized spectrum,

$$\begin{aligned} A(\bar{\nu}) &= -\log_{10}\left[\frac{\boldsymbol{I}_S(\bar{\nu})}{I_0(\bar{\nu})}\right] \\ &= \frac{4\pi\bar{\nu}k(\bar{\nu})b}{2.303} - \log_{10}\left[\frac{T_S(\bar{\nu})}{T_0(\bar{\nu})}\right] \end{aligned} \qquad (35)$$

The molar absorptivity is defined as

$$a(\bar{\nu}) = \frac{4\pi\bar{\nu}k(\bar{\nu})}{2.303\rho} \qquad (36)$$

where ρ is the concentration of the absorbing species. Finally, in the ideal case where the sample-free transfer function, $T_0(\bar{\nu})$, is equal to the transfer function with the sample, $T_S(\bar{\nu})$, Beer's law

$$A(\bar{\nu}) = a(\bar{\nu})b\rho \qquad (37)$$

can be recovered from eq 35.

With comparison of eqs 35 and 37, it may be recognized that the recorded absorbance spectrum should be corrected for optical effects to recover analytically meaningful spectra that are independent of the instrument and the sample geometry. Comparing eq 33 with the rigorous model in the previous section, it should be noted that the simple model does not fully take into account the structure of the object (beyond the path length b) nor the real part of the refractive index. These two factors are known to lead to interference and dispersion effects in bulk-sample spectroscopy.[30,31] Restated, in the simple model the transmission or reflection coefficient $T_S(\bar{\nu})$ is considered to be independent of the sample geometry and the properties of the sample and substrate. The model of eq 33 also does not account for the angle(s) of illumination and detection, this can be particularly important in microspectroscopy, as focusing results in simultaneous illumination with waves of many incidence angles. The impact of neglecting these factors on the data is apparent in the simulations that follow. The various effects leading to spectral distortions are identified and systematically quantified through the use of the developed model.

In the rest of this article, the model described is first experimentally validated using a benchmarking sample. The effects of the focusing optics of the imaging system are then isolated in simulation by considering a hypothetical idealized sample that eliminates the sample-induced distortions described by the second term in eq 35. Transmission and transflection geometries using common substrates are then simulated and it is seen that transflection measurements in particular are susceptible to sample-induced distortions. These distortions are seen to be exacerbated if a substrate of intermediate index is used. Further sample-induced distortions are predicted if an air gap is present between the substrate and the sample. Finally, the correspondence between a simplified single-ray model and the fully focused model is examined.

(30) Allara, D. L.; Baca, A.; Pryde, C. A. *Macromolecules* **1978**, *11*, 1215–1220.
(31) Zhang, Z. M. *J. Heat Transfer* **1997**, *119*, 645–647.

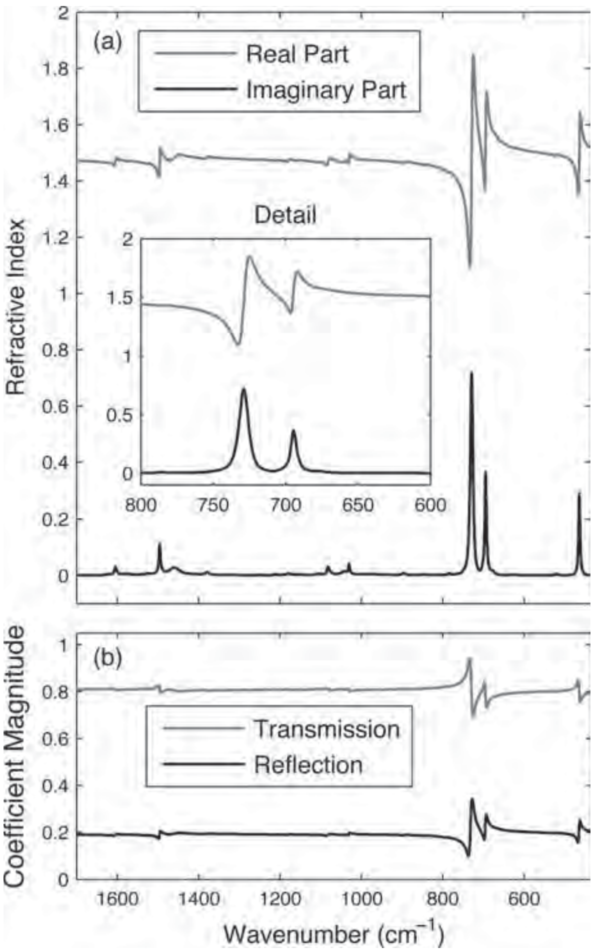


Figure 6. The (a) complex refractive index of toluene[32] and (b) the magnitude of the normal-incidence complex transmission and reflection coefficients for an air−toluene boundary. The Supporting Information includes graphs of the refractive indices of the other materials considered in this article.

To make predictions from the theoretical model, toluene is used as a homogeneous sample of interest. Toluene exhibits distinct and clearly identifiable absorption modes of varying strength over the entire mid-IR region, making it an ideal sample. In addition, toluene has a well characterized complex refractive index[32] [shown in Figure 6a] which is publicly available.[33] Refractive index changes and anomalous dispersion[34] in the vicinity of absorption peaks can be clearly observed [e.g., see inset in Figure 6a]. This variation of refractive index, in both the real and imaginary parts, affects the recorded data. A simple illustration of these effects is shown in Figure 6b, where the transmission and reflection coefficients at an air−toluene boundary can be seen to have structure influenced by the dispersive real index profile.

Experimental Validation of the Model. The model presented here was validated by performing microspectroscopy measurements on a well characterized benchmarking sample. The sample was designed such that both transmission and transflection measurements would result in significant signal. The model should then be able to accurately predict both sets of data from a single description of the sample−substrate system.

The sample was prepared by first forming a thin (≈75 nm) germanium layer by sputter coating on a barium fluoride (BaF_2) disk. A common photoresist material, SU-8 2000.5 (MicroChem Corp., Newton, MA), was spin coated to an approximate thickness of 10 μm and pattern cured by UV exposure using a standard USAF 1951 target (Edmond Optics, Barrington, NJ). Postcuring, the entire sample was baked at 95 °C and developed as per standard protocols.[35] A postbake at 150 °C for 5 min was performed to ensure complete polymerization and long-term stability.

The sample data were recorded on a Varian Stingray system using a mid-IR interferometer and microscopy with glass apertures. A narrowband, liquid nitrogen cooled detector is used to record spectra. Data are recorded at an undersampling ratio of 2 referenced to the He−Ne laser, zero-filled by a factor of 2, and transformed using Happ−Genzel apodization. Single beam spectra acquired for the sample (a position near the center of a larger region of SU-8 and far from an transverse features) and background (a position with no SU-8) are subsequently converted to absorbance spectra. Both transmission and transflection mode data were acquired without perturbing the sample.

The SU-8 polymer is the sample layer to be characterized, while the refractive indices are known for the thin germanium[36] layer and the barium fluoride[37] substrate. Background single beam spectra, Figure 7d, are recorded on a region of the sample without SU-8, and the sample single beam spectra, Figure 7a, are recorded on a region of the sample with SU-8. If absorbance calculations are performed according to eq 35, the transmission and transflection results, plotted in Figure 7c, are not consistent. Significant interference effects are visible, and peak shapes, locations, and heights can be seen to differ significantly, despite the fact that the measurements were taken from the same sample.

To correctly interpret the data, it is necessary to include optical effects, as modeled in this work. As a first step, the source spectrum, $|\mathbf{P}(s_x, s_y, \bar{\nu})|^2$ was determined from the background measurement. It is assumed that the illumination is constant across the entrance pupil so that the spectrum is not dependent on s_x and s_y. The numerical aperture of the Cassegrain and the Cassegrain obstruction were found to be best modeled as 0.40 and 0.26, respectively. The expected reflection and transmission coefficients from the air, germanium, barium fluoride, air system were calculated using the microspectroscopy model and divided out (see eq 34). The resulting transmission and transflection single beam spectra of the source are shown in Figure 7g. Since the instrument uses different optical paths for the transmission and transflection measurements, these two source spectra cannot be expected to be equal or proportional. It can, however, be seen that the source spectral profiles are qualitatively consistent, which was not the case before transmission and reflection effects were considered, see Figure 7d.

Once the source profiles are established, a preliminary estimate of absorbance can be found. Data were predicted by modeling the sample index as purely real with $n_0(\bar{\nu}) = 1.4017$. The data

[32] Bertie, J. E.; Jones, R. N.; Apelblat, Y.; Keefe, C. D. *Appl. Spectrosc.* **1994**, *48*, 127–143.

[33] http://keefelab.cbu.ca/?page_id=19.

[34] Saleh, B. E. A.; Teich, M. C. In *Fundamentals of Photonics*; Wiley-Interscience: New York, 1991; Chapter 5, pp 176−179.

[35] Processing Guidelines for SU-8 Permanent Epoxy Negative Photoresist. http://www.microchem.com/products/pdf/SU-82000DataSheet2000thru2015Ver4.pdf.

[36] Barnes, N. P.; Piltch, M. S. *J. Opt. Soc. Am.* **1979**, *69*, 178–180.

[37] Malitson, I. H. *J. Opt. Soc. Am.* **1964**, *54*, 628–632.

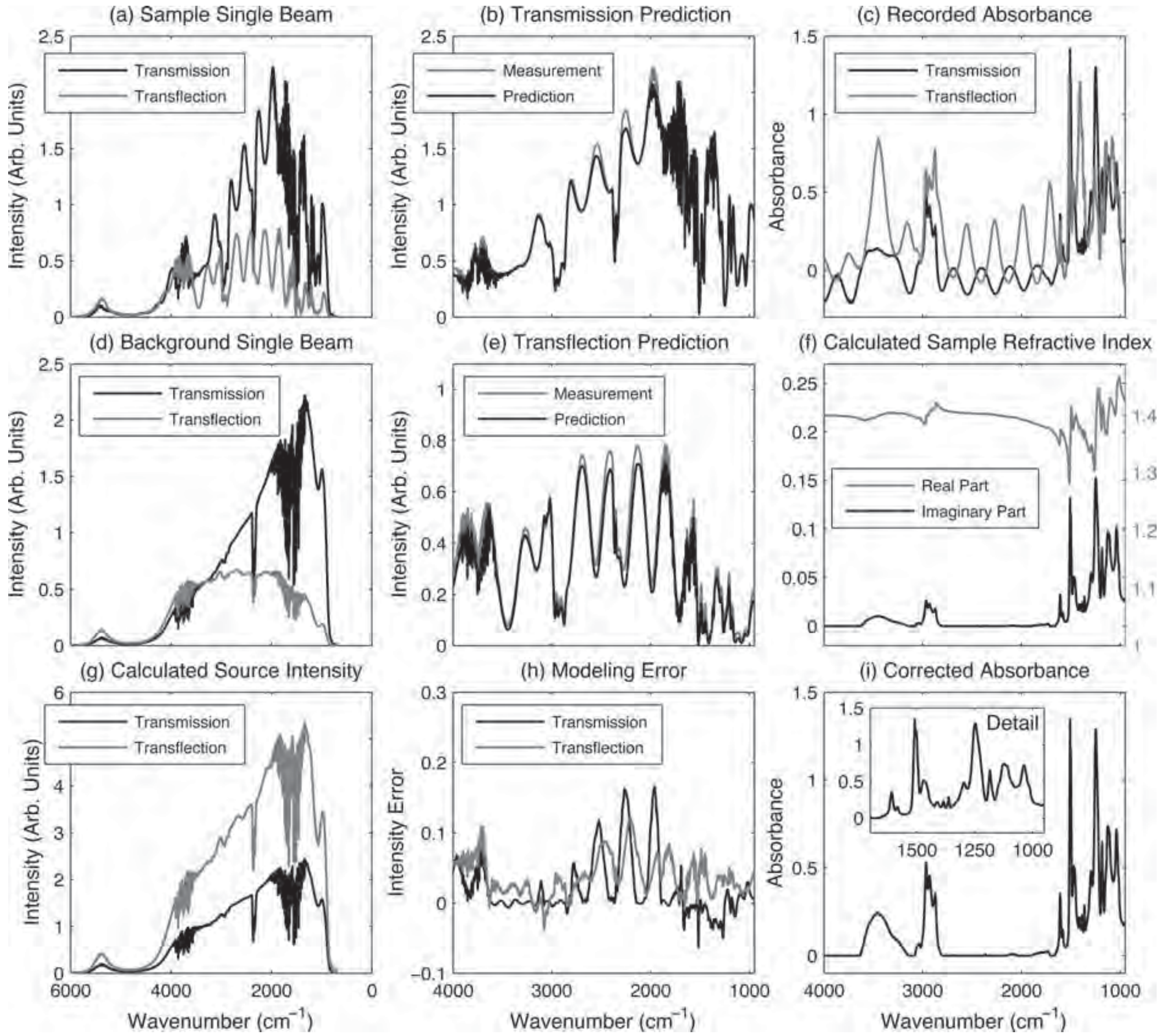


Figure 7. Experimental data and quantities used in the modeling of the benchmarking sample: (a) the measured single beam spectra for the substrate and sample (SU-8 polymer layer), (d) the measured single beam spectra for the substrate alone, (c) the absorbance as calculated from the ratios of the single beam spectra using eq 35, (g) the source spectra $|\mathbf{P}(\bar{\nu})|^2$, as calculated by compensating the background spectra for the effects of the substrate, (b) the single beam transmission spectrum predicted by the model, compared to that measured, (e) the single beam transflection spectrum predicted by the model, compared to that measured, (f) the residual differences between the plots in parts b and e, (f) the complex refractive index calculated using the microspectroscopy model and Kramers−Kronig analysis, (i) the recovered absorbance of SU-8, as calculated using the imaginary part of the refractive index shown in part f and eq 35.

predicted from the uniform index were used as an improved background measurement which captures interference effects, such as those seen in Figure 7a but not in the original background measurement of Figure 7d. With the use of eq 35, an estimate of the imaginary index $k(\bar{\nu})$ can be found from the transmission data. Applying a Kramers−Kronig calculation[38] to $k(\bar{\nu})$ gives an estimate of the variations of $n(\bar{\nu})$ about the underlying constant value $n_0(\bar{\nu})$. The estimate of the refractive index, $n(\bar{\nu}) + ik(\bar{\nu})$, is then used to predict the single beam transmission spectrum. The difference between this prediction and the measurement is then used to update the absorbance and hence the imaginary index $k(\bar{\nu})$. By iteration of the algorithmic cycle, (1) update the absorbance estimate from the difference between the measured and predicted transmission data; (2) calculate the complex refractive index from the absorbance using Kramers−Kronig analysis; (3) calculate the predicted transmission data using the model and the complex refractive index of the polymer, it is possible to converge on the complex refractive index of the polymer.[39,40]

The resulting complex index is plotted in Figure 7f and is used to predict spectra for both transmission and transflection modes. A good agreement, Figure 7b,e,h, is observed between the predicted and observed data for both transmission and transflection. The errors that are observed can likely be attributed to factors such as sample variations, unmodeled elements in the instrument optical path, and/or sample tilt. In Figure 7i the absorbance spectrum, free of optical effects, as calculated from the recorded data is shown. The agreement between predictions and measurements validates the model by demonstrating predictive power based on a physical description of the sample.

In modeling the measurement of the benchmarking sample, the refractive index of the cross-linked SU-8 layer was estimated. This estimate shows good consistency with the noncross-linked

[38] Kuzmenko, A. B. *Rev. Sci. Instrum.* **2005**, *76*, 083108.

SU-8 measurements that appear in the literature.[41,42] It should also be noted that the background value of the refractive index, $n_0(\bar{\nu})$, the exact thicknesses of the SU-8 layer (12.43 μm), and the exact thickness of the germanium layer (79 nm) were estimated by minimizing the difference between the observed and predicted data. The resultant values are consistent with the specifications used in the manufacture of the sample. It may be possible to further improve the model accuracy by including effects such as the nonuniform illumination of the Cassegrain aperture (e.g., the supports for the secondary reflector obstruct a small portion of the aperture). Nevertheless, the results presented here indicate that the level of modeling proposed here can substantially help in understanding recorded data as well as optical effects in IR microspectroscopy.

SIMULATION AND PREDICTION

Instrumentation Effects. In simulation, it is possible to separate effects due to the sample and substrate and effects due to the instrumentation. Here, this is first accomplished to investigate the dependence of the measured spectra on the numerical aperture of the imaging system. Sample-induced distortions can result from changes in reflection and/or transmission coefficients between the background and sample measurements, an effect represented in the second term of eq 35. To eliminate these coefficient changes, one can consider a nondispersive, weakly absorbing sample. Here the imaginary part of the sample index, $k(\bar{\nu})$, is taken to be 1/100 of the imaginary part of the index of toluene, and the real part of the refractive index, $n(\bar{\nu})$, is taken to be 1. Note that this is not a physically realizable material, as causality requires that changes in the absorption must necessarily be associated with changes in the real part of the refractive index.[43] However, the minimal perturbations of the complex refractive index allow the isolation of instrument-induced changes in the data.

A 200 μm thick layer of the sample is taken to be in direct contact with a substrate of barium fluoride for transmission measurements and with a substrate of gold in transflection measurements. The background measurements are simulated with only the substrate present. The large sample thickness, paired with the weak absorption, results in absorbance data comparable to those expected from an ideal measurement of a 2 μm thickness of toluene. The indices of barium fluoride and gold are calculated using published coefficients[37,44] in a Sellmeier equation. It should be noted that in this article the Sellmeier equation for the real index of barium fluoride has been extended beyond the transmission window in order to allow a consistent comparison with transflection systems for low wavenumbers. Numerical apertures of 0.2, 0.5, and 0.9 (as in Figure 4) are simulated in both transmission and transflection geometries and the illuminating light is taken to be unpolarized. Note that for each NA the central obscuration produced by the Cassegrain is taken to have a radius covering 20% of the NA. The NA of 0.5 is similar to that available in commercial microspectroscopy systems, while the NAs of 0.9 and 0.2 provide greater and lesser comparisons.

An estimate of the absorbance is found by evaluating eq 35, and the results are displayed in Figure 8a. Note that the absorbance has been normalized by the path length (in micrometers). In this idealized example, a good agreement between the measured absorbance and the actual absorbance is expected. However, overestimation of the absorbance by an amount that increases with the numerical aperture is predicted. This apparent violation of Beer's law arises because of the increasing path length through the sample associated with higher-angle rays, a phenomenon predicted by Blout et al.[11] The procedure of Blout et al. accurately predicts the errors of Figure 8a; however, a more general correction procedure must account for a coupling between measurements, sample structure, and all angles of incidence, a set of phenomena explored further in the following simulations.

Sample-Induced Distortions. Next, consider a toluene sample (i.e., with the index illustrated in Figure 6) on barium fluoride for transmission measurements and on gold for transflection measurements. To investigate the effect of sample-induced distortions, measurements are simulated for a variety of sample thicknesses. The background measurements are taken by replacing the sample with a medium of the same thickness as the sample but with index of $n = 1.47$. These background measurements are designed to represent an optimistic case, where Fabry−Perot type fringing effects in the background cancel similar effects in the sample measurement, giving a relatively good match between $T_S(\bar{\nu})$ and $T_0(\bar{\nu})$ (see eq 35). Hence, experimentally observed data will contain additional fringes arising from purely optical effects. Various methods have been proposed for correction of fringes.[45] It must be noted, however, that explicitly accounting for physical effects is likely to be more accurate than signal processing methods alone, as was shown in Figure 7. The illuminating light is taken to be unpolarized, while the NA of the system is modeled as 0.5 (with a central NA of 0.1 obscured by the secondary Cassegrain reflector).

The simulation results are shown in Figure 8b. In the transmission experiments, the estimates of absorbance are reasonably accurate. In transflection, however, errors in peak position, peak height, and band shape are predicted in the absorption spectra. Such distortions have also been observed experimentally.[46] As a consequence of the dispersion quantified by the Kramers−Kronig relations,[43] strong absorption peaks are accompanied by sharp changes in the real refractive index (e.g., see Figure 6). This results in a significant difference between the coefficients $T_0(\bar{\nu})$ and $T_S(\bar{\nu})$ seen in eq 35 and leads to distortions. Furthermore, when the sample thickness is on the scale of the wavelength, reflected and transmitted components interfere, resulting in a complicated interplay of dispersion, sample geometry, and absorption. The differences in the predicted spectra with sample thickness stem from these phenomena.

[39] Hawranek, J. P.; Jones, R. N. *Spectrochim. Acta* **1976**, *32A*, 99–109.

[40] Hawranek, J. P.; Neelakantan, P.; Young, R. P.; Jones, R. N. *Spectrochim. Acta* **1976**, *32A*, 85–98.

[41] Tan, T. L.; Wong, D.; Lee, P.; Rawat, R. S.; Patran, A. *Appl. Spectrosc.* **2004**, *58*, 1288–1294.

[42] Tan, T. L.; Wong, D.; Lee, P.; Rawat, R. S.; Springham, S.; Patran, A. *Thin Solid Films* **2006**, *504*, 113–116.

[43] Toll, J. S. *Phys. Rev.* **1956**, *104*, 1760–1770.

[44] Ordal, M. A.; Long, L. L.; Bell, R. J.; Bell, S. E.; Bell, R. R.; Alexander, R. W., Jr.; Ward, C. A. *Appl. Opt.* **1983**, *22*, 1099–1120.

[45] Griffiths, P. R.; de Haseth, J. A. In *Fourier Transform Infrared Spectrometry*, 2nd ed.; Wiley-Interscience: Hoboken, NJ, 2007; Chapter 11.1.3, pp 253−255.

[46] Gunde, M. K.; Aleksandrov, B. *Appl. Spectrosc.* **1990**, *44*, 970–974.

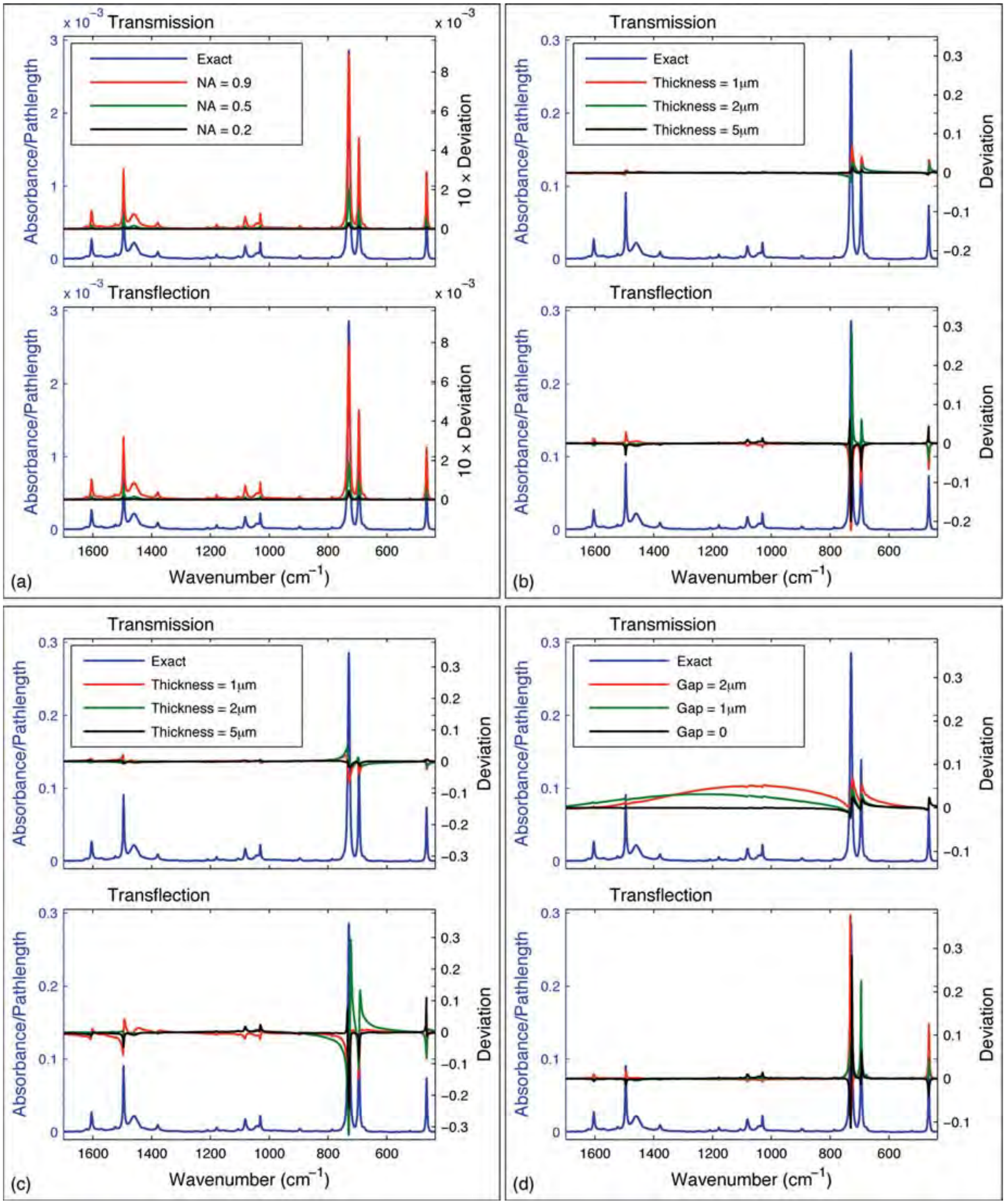


Figure 8. (a) Predicted absorbance for an idealized thick, low-absorption sample, normalized by the path length (in micrometers). Data are plotted for both transmission and transflection modalities (and for a range of numerical apertures) as differences from the ideal absorbance profile. In transmission, the substrate is barium fluoride, and in transflection, the substrate is gold. (b) Predicted path-length−normalized absorbance deviations for a toluene sample and a range of sample thicknesses. In transmission, the substrate is barium fluoride, and in transflection, the substrate is gold. (c) Predicted path-length−normalized absorbance deviations for a toluene sample and a range of sample thicknesses. In both transmission and transflection, the substrate is germanium. (d) Predicted path-length−normalized absorbance deviations when there is an air gap between the sample and the substrate. The air gap thickness is varied, while in all cases the sample thickness is 2 μm. In transmission, the substrate is barium fluoride, and in transflection, the substrate is gold. The absorbance spectra used to calculate the deviations shown here are plotted in the Supporting Information.

Sample-Induced Distortions for Substrates of Intermediate Index. The appearance of the dispersion profile (the real part of the refractive index) in absorption microspectroscopy measurements has been described.[13−15] It was noted that the estimate is more susceptible to this consequence of dispersion in transflection mode or when, for example, a higher-index substrate is used in transmission. The dispersion influence can be explained, at least in part, by effects such as those predicted in Figure 8a. Romeo and Diem[13] also observed a similar feature at sample edges; this phenomenon is investigated in the follow-up article.[22]

If both the transmission and transflection substrates are germanium (with background measurements taken on the bare germanium), the spectra predicted are shown in Figure 8c. The refractive index of germanium was calculated using published coefficients[36] for a Sellmeier model, and all other simulation parameters are the same as for Figure 8b. The germanium substrate can be seen to give seemingly confounding results; the transflection spectra have severe distortions including negative values of absorbance that are not physically realizable, while the transmission spectra have distorted band shapes and amplitudes. Hence, it is clear that the high index of germanium makes it unsuitable for accurate transmission or transflection measurements—without corrections for optical distortions. An ideal transmission substrate has an index matched to the sample, while an ideal transflection substrate is, for example, a strong conductor. At the toluene−germanium boundary both the transmission and reflection coefficients are significant and both are relatively sensitive to the sample index. As a result, the real part of the index is strongly coupled into the measurement. This coupling is particularly noticeable in the transflection measurement and results in apparently negative absorbance measurements. The transflection simulations of Figure 8b,c illustrate how the presented framework can be used to examine spectral distortions introduced in the transflection modality and suggests how explicit optical modeling may be useful in the design of transflection substrates.

Air-Gap-Induced Distortions. It is not uncommon in the mounting of a sample on the substrate to introduce a small air gap between the two; alternatively, the sample itself may contain a void. In Figure 8d, spectral distortions caused by such voids are shown for a variety of air gaps. The sample material is again toluene, and the background measurements are taken without accounting for the gap. Significant changes in peak shape, amplitude, and position can again be seen in transflection. The distortions are less severe in transmission although a significant nonzero baseline is observed. Findings consistent with these simulations have been reported with an underlying Matrigel layer and observed to depend on layer thickness[14] as seen here. However the effect was in that work attributed to a scattering effect based on a qualitative analysis. An alternative qualitative analysis attributes distortions to contributions from reflections from the top surface of a sample.[15] It was also reported by Romeo and Diem that poorly adhered or thin samples may produce a dispersive line shape,[13] consistent with results shown in Figure 8d. The rigorous model developed here accounts for both the observed results in a quantitative manner, as well as acting as a guide to understand potentially confounding effects in sample preparation. An understanding of this effect is especially relevant to cytological analyses in which single cells are analyzed for malignancy. Sample preparation becomes critical in those applications and has been reported to be a major challenge in developing IR microscopy for cytology.[47] The effect on tissue samples can be expected to be less drastic, as individual cell spectra are usually less important within the greater tissue structure, and both the spectral and spatial organization of the cells can be employed for effective diagnoses.[48]

Comparison with Bulk (Macro) Spectroscopy. The simulations presented above have shown how sample structure and the real (dispersive) part of the refractive index affect the recorded spectral data. These effects produce apparent deviations from Beer's law if the simple model of eq 33 is applied. The importance of optical effects has been recognized for some time,[49−51] particularly in reflection-based modalities, and algorithms[38,52,53] have been developed to calculate the complex refractive index from certain types of data measured in bulk spectroscopy. In systems without tight focusing, this type of approach has been applied to correct for the apparent artifacts[21,39,40,54−59] and should be used where possible. In addition to general interference and dispersion effects (as observed without tight focusing in bulk-sample spectroscopy), the model developed in this work takes into account optical effects produced by the tightly focused illumination and collection of light. If focusing effects are negligible in comparison with the effects already modeled in bulk spectroscopy, it can be expected that existing correction algorithms will interpret microspectroscopy data correctly.

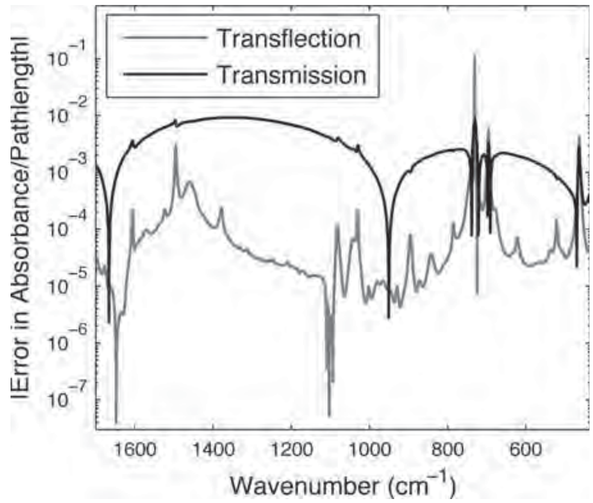


Figure 9. Magnitude of the difference between the data predicted in Figure 8d (for an air gap of 2 μm thickness) and the data predicted using a comparable single ray model.

Figure 9 shows the difference between the focused-model, 2 μm-air-gap data of Figure 8d and those calculated for the same sample but using a single representative ray path (i.e., a model without focusing). For the transmission system, the representative ray path is taken to be at normal incidence, and for the transflection system, the median reflected path is chosen. The single-ray approach does not capture effects due to focused path length difference [as illustrated in Figure 8a] and, as seen in Figure 9, will not fully capture the behavior of the tightly focused system. For this example, the

(47) Romeo, M.; Mohlenhoff, B.; Diem, M. *Vib. Spectrosc.* **2006**, *42*, 9–14.
(48) Pounder, F. N.; Bhargava, R. Submitted for publication.

(49) Greenler, R. G. *J. Chem. Phys.* **1966**, *44*, 310–315.
(50) Greenler, R. G. *J. Chem. Phys.* **1969**, *50*, 1963–1968.
(51) Mendelsohn, R.; Brauner, J. W.; Gericke, A. *Annu. Rev. Phys. Chem.* **1995**, *46*, 305–334.
(52) Dienstfrey, A.; Greengard, L. *Inverse Probl.* **2001**, *17*, 1307–1320.
(53) De Sousa Meneses, D.; Rousseau, B.; Echegut, P.; Simon, P. *Appl. Spectrosc.* **2007**, *61*, 1390–1397.
(54) Andermann, G.; Caron, A.; Dows, D. A. *J. Opt. Soc. Am.* **1965**, *55*, 1210–1216.
(55) Hawranek, J. P.; Neelakantan, P.; Young, R. P.; Jones, R. N. *Spectrochim. Acta* **1976**, *32A*, 75–84.
(56) Bertie, J. E.; Apelblat, Y. *Appl. Spectrosc.* **1996**, *50*, 1039–1046.
(57) Yamamoto, K.; Ishida, H. *Vib. Spectrosc.* **1997**, *15*, 27–36.
(58) MacDonald, S. A.; Schardt, C. R.; Masiello, D. J.; Simmons, J. H. *J. Non-Cryst. Solids* **2000**, *275*, 72–82.
(59) Moore, D. S.; McGrane, S. D.; Funk, D. J. *Appl. Spectrosc.* **2004**, *58*, 491–498.

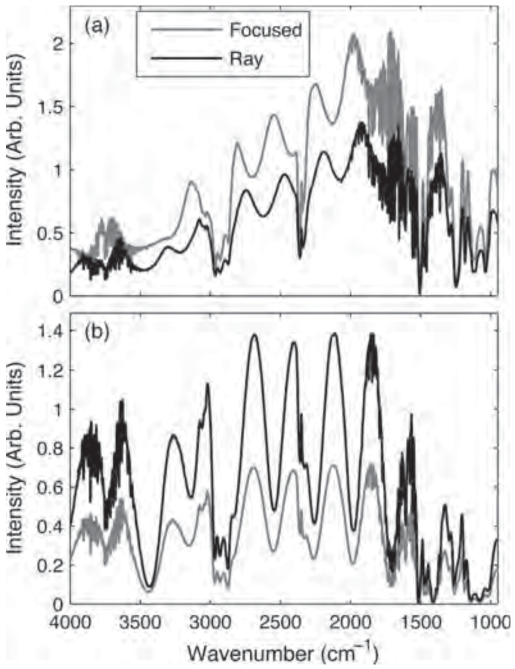


Figure 10. Ray-based and fully focused predictions for (a) transmission and (b) transflection modalities and the experimental benchmarking sample.

single-ray simplification produces errors in the transflection system in particular. It is noteworthy that contemporary instruments can produce signals with low enough noise to observe absorbance values in the 10^{-1} to 10^{-4} range. Hence, these errors are significant, and the detailed model developed here should be used.

Differences between the focused and single-ray models arise from the angular dependence of the light−sample interaction. The difference between the full model spectrum and that predicted using a single ray model is shown in Figure 9. It is seen that this angular dependence can be significant, particularly around regions of high absorption. For the simpler samples considered in Figure 8b (i.e., air−sample−substrate systems with no air gap), the angular dependence is less critical and gives maximum path length-normalized absorbance errors of 0.0044 in transmission and 0.011 in transflection, as compared to maximum errors of 0.0046 and 0.11 in transmission and transflection, respectively, in Figure 9. Conversely, many-layer sample−substrate systems, with comparable transmission and reflection coefficients at layer boundaries, may be highly sensitive to incidence angle and hence to focusing effects. In Figure 10, a comparison between ray-based and focused models for the experimental benchmarking system (see Figure 7) is presented. For this more complicated sample−substrate system, the use of the ray-based model introduces significant errors. Focusing effects therefore play a very significant role when modeling the benchmarking sample used in the experimental validation of the model.

The analyses presented here should be used as a guide to estimate the precision in the data. The first implication is that the choice of sampling mode and/or substrate greatly influences the magnitude and form of systematic error introduced into the measurement. A second result demonstrates that there is a dramatic difference in the precision achievable by transmission mode and transflection mode microspectroscopy. The distortion is nonlinear and not trivial to correct. One practical implication is that the noise in the data acquired must be no smaller than the observed deviation from the true spectrum. Any further reduction in noise would make the analytical conclusions limited by systematic distortions and not random noise. In general, the presented theoretical framework should be considered a starting point for detailed optical modeling in specific studies. In biomedical applications, where spectral assignments are challenging and spectral changes are small, detailed modeling can be expected to be important in understanding biochemical changes accurately.

CONCLUDING REMARKS

A mathematical model for mid-IR microspectroscopy has been derived by solving Maxwell's equations in layered media and for focused illumination and detection. Predictions given by this model are consistent with experimental results and with observations reported in the literature. It is seen that the interplay of focusing, the sample geometry, and strong dispersion fully accounts for the spectral response and apparent artifacts for simple homogeneous systems. Additional spectral effects that are produced by scattering within heterogeneous materials are addressed in part II of this work.

The model developed here can be applied to both transflection and transflection collection geometries. While transmission spectra demonstrated some robustness to distortions, transflection systems were seen to be particularly sensitive to focusing, dispersion, and sample−structure induced distortions. Ideally the distortions observed may be corrected by mathematically inverting the developed model, in order to estimate optical constants of the sample directly. However, in many cases of interest, the sample structure (i.e., the materials present in sample layers and the layer thicknesses) may not be known. This complicates the inversion process, as the sample geometry must be coestimated with the optical constants of the material of interest.

Spectral distortions due to sample structure (e.g., interference between interfaces) and dispersion have previously been reported for systems that do not employ tight focusing. The model presented here describes tightly focused fields throughout the sample and also predicts focusing dependent distortions that may impact the measured spectra for certain sample geometries. In comparison to typical experimental noise in modern IR microspectroscopy systems, the effects were found to be significant. Consequently, the model described provides a means to understand distortions that may limit the analytical capability of IR microspectroscopy.

ACKNOWLEDGMENT

This work was supported in part by the Grainger Foundation for Emerging Technologies, the Department of Defense Prostate Cancer Research Program, and Susan G. Komen for the Cure. The authors would also like to thank Dr. Gokulakrishnan Srinivasan for fabricating the SU-8 sample and Dr. Michael J. Walsh for help in acquiring data.

SUPPORTING INFORMATION AVAILABLE

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review September 15, 2009. Accepted December 31, 2009.

AC902067P

Theory of Mid-infrared Absorption Microspectroscopy: II. Heterogeneous Samples

Brynmor J. Davis,[†] P. Scott Carney,[‡] and Rohit Bhargava[*,†]

Department of Bioengineering, Department of Electrical and Computer Engineering, and the Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana−Champaign, Urbana, Illinois 61801

Fourier transform infrared (FT-IR) spectroscopic imaging combines the specificity of optical microscopy with the spectral selectivity of vibrational spectroscopy. There is increasing recognition that the recorded data may be dependent on the optical configuration and sample morphology in addition to its local material spectral response, but a quantitative framework for predicting such dependence is lacking. Here, a theory is developed to relate recorded data to the spectral and physical properties of heterogeneous samples. The modeling approach combines optical theory through rigorous coupled wave analysis with modeling of sampling geometry and sample structure. The interplay of morphology and dispersion are systematically explored using increasingly sophisticated samples to illustrate the dependence of the detected optical intensity on the spatial sample structure. Predictions of spectral distortions arising from the sample structure are quantified, and experimental validation of the developed theory is performed using a microfabricated standard from a commercial photoresist polymer. The developed framework forms a basis for understanding sample induced distortions in spectroscopic IR microscopy and imaging.

Fourier transform infrared (FT-IR) spectroscopic imaging is a rapidly emerging technology that combines the spatial specificity of optical microscopy with the chemical selectivity of vibrational spectroscopy.[1−4] It is commonly misconceived that FT-IR imaging is a simple extension of conventional infrared spectroscopy using a different sampling accessory, namely a microscope. From the optics perspective, similarly, it is tempting to conclude that FT-IR imaging is an extension of optical microscopy with discrimination of IR light by wavelength. In this series of articles, it is shown that neither characterization is accurate. In the previous article,[5] optical theory for IR microscopy was developed and it was demonstrated that the combination of the sample−substrate structure and optical configuration can result in significant distortions in data recorded from homogeneous samples. Briefly, optical theory was applied to model interrogation of a sample that was assumed to consist of a homogeneous layer in a sample−substrate structure with no transverse variation. The sample was characterized by upper and lower boundaries and by the frequency dependent complex relative permittivity $\varepsilon(\bar{\nu})$, or equivalently, by a constant complex refractive index.

In this article, the analysis is extended to heterogeneous samples that vary in the lateral sample plane. The sample is characterized by upper and lower boundaries as well as a transverse structure defined by permittivity, $\varepsilon(x, y, \bar{\nu})$. An example of this type of structure is shown in Figure 1. While the sample has nontrivial structure in the imaging plane, it is assumed to be piecewise constant as a function of depth. Such a model is appropriate for thin samples as are usually encountered in IR microspectroscopy. This structure is amenable to analysis through coupled wave theory. The notation used here is consistent with the first article[5] and is also listed in the glossary of Table S1 in the Supporting Information.

In the earliest studies,[6] it was noted that heterogeneous sample structure distorts both the apparent spectrum and the apparent spatial structure in FT-IR imaging. Other authors have also attributed spectral distortions to heterogeneous sample structure.[7−9] While experiments[10] demonstrated that distortions arose from a mismatch of refractive index between domains in the sample, a complete theoretical model to predict the effects of heterogeneous samples on observed spectra and spatial structure has not been presented. The absence of such a model can lead to misinterpretation of spatial structure and/or spectral changes observed at the boundaries of domains. The full analytical capability of FT-IR imaging can only be realized through proper modeling of the optical physics of the combined sample-instrument system. These models help, first, to understand the true spectral and structural content of the data. Second, they help provide measures of the systematic error due to distortions. Studies that claim chemical

[*] To whom correspondence should be addressed. E-mail: rxb@illinois.edu.

[†] Department of Bioengineering and the Beckman Institute.

[‡] Department of Electrical and Computer Engineering and the Beckman Institute.

(1) *Spectrochemical Analysis Using Infrared Multichannel Detectors*; Bhargava, R., Levin, I. W., Eds.; Sheffield Analytical Chemistry Series; Wiley-Blackwell: Oxford, U.K., 2005.

(2) Colarusso, P.; Kidder, L. H.; Levin, I. W.; Fraser, J. C.; Arens, J. F.; Lewis, E. N. *Appl. Spectrosc.* **1998**, *52*, 106A–120A.

(3) Levin, I. W.; Bhargava, R. *Annu. Rev. Phys. Chem.* **2005**, *56*, 429–474.

(4) Lewis, E. N.; Treado, P. J.; Reeder, R. C.; Story, G. M.; Dowrey, A. E.; Marcott, C.; Levin, I. W. *Anal. Chem.* **1995**, *67*, 3377–3381.

(5) Davis, B. J.; Carney, P. S.; Bhargava, R. *Anal. Chem.* DOI: 10.1021/ac902067p.

(6) Bhargava, R.; Wang, S.-Q.; Koenig, J. L. *Macromolecules* **1999**, *32*, 8989–8995.

(7) Budevska, B. O. *Vib. Spectrosc.* **2000**, *24*, 37–45.

(8) Mohlenhoff, B.; Romeo, M.; Diem, M.; Wood, B. R. *Biophys. J.* **2005**, *88*, 3635–3640.

(9) Bassan, P.; Byrne, H. J.; Bonnier, F.; Lee, J.; Dumas, P.; Gardner, P. *Analyst* **2009**, *134*, 1586–1593.

(10) Bhargava, R.; Wang, S.-Q.; Koenig, J. L. *Appl. Spectrosc.* **1998**, *52*, 323–328.

10.1021/ac902068e © 2010 American Chemical Society
Published on Web 04/14/2010

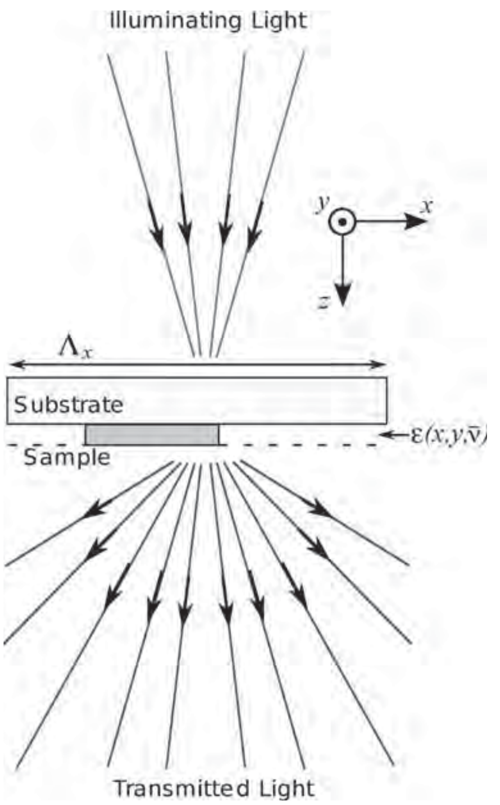


Figure 1. An illustration of the type of sample and substrate geometry considered in this article. Here the sample (a slab of finite extent) is illuminated through a substrate. The sample layer is defined by the permittivity $\varepsilon(x, y, \bar{\nu})$, and the region of interest is of width Λ_x in the x direction. In contrast to the previous article, the sample may scatter light outside of the illumination angles.

or structural changes at edges of domains might employ the model reported here to verify that the magnitude of those changes is indeed larger than those due to optical effects alone. In this article, optical theory for analysis of heterogeneous structures in mid-IR imaging is developed. The variation of certain parameters in the model is predicted to lead to specific distortions. Predictions are compared to experimental data.

THEORETICAL MODEL

In the preceding work,[5] it was shown that each planewave mode of the electric field (indexed by the propagation directions s_x and s_y) may be propagated through the sample−substrate system independently. When transverse sample structure is introduced this is no longer true. Optical effects such as scattering and refraction induce coupling between the modes. These effects are calculated below using rigorous coupled wave analysis,[11−13] which was originally developed for modeling diffraction gratings. While there are alternative methods which could be used to solve the problem at hand,[14−16] coupled wave analysis provides a clear description of how the transverse structure of the object couples planewave modes. The coupled wave method is also widely used, and the associated numerical implementation is well studied.[17] In the following presentation, rigorous coupled wave analysis is briefly described and applied to the mid-IR imaging problem in order to explain artifacts, for example, from edge scattering.

It is assumed that the transverse area of interest in the sample is some finite range $\Lambda_x \times \Lambda_y$ in Cartesian coordinates x, y. The object within this range can then be represented in the Fourier series

$$\varepsilon(x,y,\bar{\nu}) \approx \sum_{p=-N_U}^{N_U-1} \sum_{q=-N_W}^{N_W-1} \phi_{p,q}(\bar{\nu}) \exp[i(pUx + qWy)] \tag{1}$$

where $U = 2\pi/\Lambda_x$ and $W = 2\pi/\Lambda_y$. The Fourier series has been truncated to $2N_U$ terms in the x direction and by $2N_W$ terms in the y direction. Note that this representation repeats the object periodically outside the region of interest. However, the problem is formulated below so that light is focused into the single period $\Lambda_x \times \Lambda_y$ with negligible intensity outside this area.

The reciprocal of the permittivity is well-defined and will be useful in the analysis below. This function can also be represented as the Fourier series,

$$[\varepsilon(x,y,\bar{\nu})]^{-1} \approx \sum_{p=-N_U}^{N_U-1} \sum_{q=-N_W}^{N_W-1} \psi_{p,q}(\bar{\nu}) \exp[i(pUx + qWy)] \tag{2}$$

As with the first article,[5] the incident field is decomposed into a collection of constituent planewaves. Each individual planewave component is infinite in extent and thus impinges on the periodic extension of the sample structure. A localized response is generated by summing over the planewave spectra near the end of the calculation, but for intermediate steps, it is useful to be able to appeal to the formal periodicity.

Consider an incident planewave with Cartesian transverse spatial frequency components δ and σ, that is, a field proportional to $\exp[i(\delta x + \sigma y)]$ in a fixed z plane. The spatial periodicity of the sample implies that the scattered field consists only of planewave components with transverse spatial frequencies that are shifted from those of the incident field by integer multiples of the constants, U and W. Explicitly,

$$u_p = pU + \delta \tag{3}$$

$$w_q = qW + \sigma \tag{4}$$

That is, through interacting with the sample, an incident planewave with transverse frequencies δ and σ must give rise to planewaves with transverse dependence of the form $\exp[i(u_px + w_qy)]$, due to the translational periodicity of the problem. At $p = q = 0$, the undiffracted component is obtained and all other values represent diffracted modes.

For reasons similar to those given above, the field in any fixed-z plane of the sample must be composed of fields with the same

[11] Moharam, M. G.; Gaylord, T. K. *J. Opt. Soc. Am.* **1981**, *71*, 811–818.
[12] Moharam, M. G.; Gaylord, T. K. *J. Opt. Soc. Am.* **1983**, *73*, 1105–1112.
[13] Gaylord, T. K.; Moharam, M. G. *Proc. IEEE* **1985**, *73*, 894–937.
[14] Li, L. *J. Opt. Soc. Am. A* **1996**, *13*, 1024–1035.
[15] Li, L. *J. Opt. Soc. Am. A* **1997**, *14*, 2758–2767.
[16] Jin, J. *The Finite Element Method in Electromagnetics*, 2nd ed.; Wiley-IEEE Press: New York, 2002.
[17] Moharam, M. G.; Grann, E. B.; Pommet, D. A.; Gaylord, T. K. *J. Opt. Soc. Am. A* **1995**, *12*, 1068–1076.

transverse frequencies given in eqs 3 and 4. Therefore, in the sample layer (indexed by layer $\ell = \Delta$), between $z^{(\Delta-1)}$ and $z^{(\Delta)}$, the electric field vector can be written in the form

$$\mathbf{E}^{(\Delta)}(x, y, z, \bar{\nu}) = \sum_p \sum_q \begin{bmatrix} X_{p,q}(z, \bar{\nu}) \\ Y_{p,q}(z, \bar{\nu}) \\ Z_{p,q}(z, \bar{\nu}) \end{bmatrix} \exp[i(u_p x + w_q y)] \tag{5}$$

Note that while the Fourier transform of the object function was truncated in eq 1, the field resulting from scattering from this approximation to the object need not be similarly band-limited. However, it is necessary in the numerical calculation of $\mathbf{E}^{(\Delta)}(x, y, z, \bar{\nu})$ to make a potentially different truncation of eq 5. This truncation is made such that the diffracted-field coefficients $X_{p,\,q}(z, \bar{\nu})$, $Y_{p,\,q}(z, \bar{\nu})$, and $Z_{p,\,q}(z, \bar{\nu})$ have decayed to a negligible level before the truncation point.

In the inhomogeneous sample layer, the magnetic field is nontrivially related to the electric field (c.f., the relationship in a homogeneous layer[18]). Hence it will be convenient to describe the magnetic field separately as

$$\mathbf{H}^{(\Delta)}(x, y, z, \bar{\nu}) = \sqrt{\frac{\varepsilon_0}{\mu_0}} \sum_p \sum_q \begin{bmatrix} I_{p,q}(z, \bar{\nu}) \\ J_{p,q}(z, \bar{\nu}) \\ K_{p,q}(z, \bar{\nu}) \end{bmatrix} \exp[i(u_p x + w_q y)] \tag{6}$$

In the homogeneous layers, e.g., in the substrate or in the homogeneous sample addressed in the preceding article, each component of the planewave spectrum of the field can be propagated independently and the results can be summed to find the field at any given plane.[19] In the structured sample considered here, the relationship between fields in distinct transverse planes is more complicated and this is reflected in the very general dependence of eqs 5 and 6 on z. The evolution of the electric and magnetic fields with z is found using the Maxwell–Faraday equation and Ampère's circuital law. With the use of the time harmonic form and the fact that $c = 1/(\varepsilon_0\mu_0)^{1/2}$, these can be written

$$\nabla \times \mathbf{E}(\mathbf{r}, \bar{\nu}) = i2\pi\bar{\nu}\sqrt{\frac{\mu_0}{\varepsilon_0}}\mathbf{H}(\mathbf{r}, \bar{\nu}) \tag{7}$$

$$\nabla \times \mathbf{H}(\mathbf{r}, \bar{\nu}) = -i2\pi\bar{\nu}\varepsilon(\mathbf{r}, \bar{\nu})\sqrt{\frac{\varepsilon_0}{\mu_0}}\mathbf{E}(\mathbf{r}, \bar{\nu}) \tag{8}$$

Substituting eqs 5 and 6 into eq 7 and equating coefficients for each transverse frequency pair (u_p, w_q) results in the equations

$$\frac{\mathrm{d}X_{p,q}(z, \bar{\nu})}{\mathrm{d}z} = i2\pi\bar{\nu}J_{p,q}(z, \bar{\nu}) + iu_p Z_{p,q}(z, \bar{\nu}) \tag{9}$$

$$\frac{\mathrm{d}Y_{p,q}(z, \bar{\nu})}{\mathrm{d}z} = -i2\pi\bar{\nu}I_{p,q}(z, \bar{\nu}) + iw_q Z_{p,q}(z, \bar{\nu}) \tag{10}$$

$$K_{p,q}(z, \bar{\nu}) = \frac{1}{2\pi\bar{\nu}}[u_p Y_{p,q}(z, \bar{\nu}) - w_q X_{p,q}(z, \bar{\nu})] \tag{11}$$

Substituting eqs 1, 5, and 6 into eq 8 and equating transverse frequency pairs for the x and y components of the vector equation gives the equations

$$\frac{\mathrm{d}I_{p,q}(z, \bar{\nu})}{\mathrm{d}z} = -i2\pi\bar{\nu}\sum_{p''}\sum_{q''}\phi_{p-p'',q-q''}(\bar{\nu})Y_{p'',q''}(z, \bar{\nu}) + iu_p K_{p,q}(z, \bar{\nu}) \tag{12}$$

$$\frac{\mathrm{d}J_{p,q}(z, \bar{\nu})}{\mathrm{d}z} = i2\pi\bar{\nu}\sum_{p''}\sum_{q''}\phi_{p-p'',q-q''}(\bar{\nu})X_{p'',q''}(z, \bar{\nu}) + iw_q K_{p,q}(z, \bar{\nu}) \tag{13}$$

Equating transverse frequency pairs, the z component in eq 8 can be found by first dividing both sides of the equation by $\varepsilon(\mathbf{r}, \bar{\nu})$. The expression for the reciprocal of $\varepsilon(\mathbf{r}, \bar{\nu})$, eq 2, can then be used to give

$$Z_{p,q}(z, \bar{\nu}) = -\frac{1}{2\pi\bar{\nu}}\left\{\sum_{p''}\sum_{q''}\psi_{p-p'',q-q''}(\bar{\nu})[u_{p''}J_{p'',q''}(z, \bar{\nu}) - w_{q''}I_{p'',q''}(z, \bar{\nu})]\right\} \tag{14}$$

The results seen in eqs 9–14 determine how the electric and magnetic fields propagate through the sample layer. The dependence on $K_{p,\,q}(z, \bar{\nu})$ and $Z_{p,\,q}(z, \bar{\nu})$ can be eliminated by substituting eq 14 into eqs 9 and 10 and eq 11 into eqs 12 and 13. The result is four sets of coupled first-order differential equations. The (u_p, w_q) frequency pairs retained in eqs 5 and 6 are then placed in a one-dimensional ordering, indexed by m. Using this one-dimensional ordering, each set of functions can be arranged as a $N_F \times 1$ column vector, where N_F is the number of terms retained. These vectors can then be concatenated and the system of differential equations written in the form

$$\begin{bmatrix} \frac{\mathrm{d}\mathbf{X}(z, \bar{\nu})}{\mathrm{d}z} \\ \frac{\mathrm{d}\mathbf{Y}(z, \bar{\nu})}{\mathrm{d}z} \\ \frac{\mathrm{d}\mathbf{I}(z, \bar{\nu})}{\mathrm{d}z} \\ \frac{\mathrm{d}\mathbf{J}(z, \bar{\nu})}{\mathrm{d}z} \end{bmatrix} = i2\pi\bar{\nu}\Phi(\bar{\nu})\begin{bmatrix} \mathbf{X}(z, \bar{\nu}) \\ \mathbf{Y}(z, \bar{\nu}) \\ \mathbf{I}(z, \bar{\nu}) \\ \mathbf{J}(z, \bar{\nu}) \end{bmatrix} \tag{15}$$

where $\Phi(\bar{\nu})$ is a $4N_F \times 4N_F$ matrix. For convenience, the dependence of Φ on $\bar{\nu}$ is suppressed for the remainder of this work.

The form[20] of Φ guarantees that eigenvalues come in pairs of opposite sign, i.e., the eigenvalues of Φ can be denoted by $\pm\gamma_1, \pm\gamma_2, \ldots, \pm\gamma_{2N_F}$. The eigenvectors of Φ are $\mathbf{g}_1, \mathbf{h}_1, \mathbf{g}_2, \mathbf{h}_2, \ldots, \mathbf{g}_{2N_F}, \mathbf{h}_{2N_F}$, where the vector $\mathbf{g}_j$ is associated with the eigenvalue γ_j and the vector $\mathbf{h}_j$ is associated with $-\gamma_j$. The eigenvalue γ_j is taken to lie in the upper half of the complex plane, that is γ_j is chosen such that its imaginary

[18] Equation 2 in ref 5.

[19] Equation 4 in ref 5.

[20] Equation 57 in ref 17.

part is positive. Note that for purely real eigenvalues, $+\gamma_j$ will be chosen to be positive.

Finding the eigenvalues and eigenvectors of Φ allows the matrix to be decomposed in the form

$$\Phi = G\Gamma G^{-1} \tag{16}$$

where Γ contains the eigenvalues on the diagonal and is zero elsewhere, and the vectors $\mathbf{g}_j$ and $\mathbf{h}_j$ are organized to form the corresponding columns of G.

An uncoupled set of $4N_F$ first order differential equations can be written as a single matrix equation $\mathrm{d}\mathbf{V}(z)/\mathrm{d}z = i2\pi\bar{\nu}\Gamma\mathbf{V}(z)$, where $\mathbf{V}(z)$ is a $4N_F \times 1$ vector. Such a set of equations is easily solved (each equation can be solved individually) and the result used to construct a solution of eq 15. That solution can be constructed as $G\mathbf{V}(z)$ so that

$$X_m(z,\bar{\nu}) = \sum_{j=1}^{2N_F} \{\beta_j g_{j,m} \exp[i2\pi\bar{\nu}\gamma_j(z - z^{(\Delta-1)})] + \hat{\beta}_j h_{j,m} \exp[-i2\pi\bar{\nu}\gamma_j(z - z^{(\Delta)})]\} \tag{17}$$

$$Y_m(z,\bar{\nu}) = \sum_{j=1}^{2N_F} \{\beta_j g_{j,m+N_F} \exp[i2\pi\bar{\nu}\gamma_j(z - z^{(\Delta-1)})] + \hat{\beta}_j h_{j,m+N_F} \exp[-i2\pi\bar{\nu}\gamma_j(z - z^{(\Delta)})]\} \tag{18}$$

$$I_m(z,\bar{\nu}) = \sum_{j=1}^{2N_F} \{\beta_j g_{\,j,m+2N_F} \exp[i2\pi\bar{\nu}\gamma_j(z - z^{(\Delta-1)})] + \hat{\beta}_j h_{j,m+2N_F} \exp[-i2\pi\bar{\nu}\gamma_j(z - z^{(\Delta)})]\} \tag{19}$$

$$J_m(z,\bar{\nu}) = \sum_{j=1}^{2N_F} \{\beta_j g_{j,m+3N_F} \exp[i2\pi\bar{\nu}\gamma_j(z - z^{(\Delta-1)})] + \hat{\beta}_j h_{j,m+3N_F} \exp[-i2\pi\bar{\nu}\gamma_j(z - z^{(\Delta)})]\} \tag{20}$$

where $g_{j,\,m}$ is the mth element of the vector $\mathbf{g}_j$, $h_{j,\,m}$ is the mth element of the vector $\mathbf{h}_j$, and β_j and $\hat{\beta}_j$ are, as yet undetermined, coefficients. The field in the sample layer is determined by eqs 17–20, with the z-polarized components given by eqs 11 and 14. The sample structure determines the values for γ, $g_{j,m}$, and $h_{j,m}$ through the eigenvalue decomposition of Φ. The $4N_F$ remaining coefficients ($2N_F\beta_j$ coefficients and $2N_F\hat{\beta}_j$ coefficients) are set by boundary conditions, i.e., the illuminating field determines these values.

A representation of the field in the homogeneous layers (e.g., the air surrounding the substrate and sample and the substrate) has been described elsewhere[19] and can be rewritten as

$$\mathbf{E}^{(\ell)}(x,y,z,\bar{\nu}) = \bar{\nu}\sum_{m=1}^{N_F} \{\mathbf{B}^{(\ell)}(m,\bar{\nu}) \exp[i2\pi\bar{\nu}s_z^{(\ell)}(m,\bar{\nu})(z - z^{(\ell-1)})] + \hat{\mathbf{B}}^{(\ell)}(m,\bar{\nu}) \exp[-i2\pi\bar{\nu}s_z^{(\ell)}(m,\bar{\nu})(z - z^{(\ell)})]\} \times \exp[i(u_{p(m)}x + w_{q(m)}y)] \tag{21}$$

The modes of the field in the homogeneous layers are here indexed by m, whereas in the previous article[5] they were indexed by the transverse propagation quantities s_x and s_y. Writing the modes in the manner above allows the field in the homogeneous layers to be matched to the field in the sample. The relationship between m and s_x and s_y is

$$s_x(m,\bar{\nu}) = \frac{u_{p(m)}}{2\pi\bar{\nu}} = \frac{p(m)}{\Lambda_x\bar{\nu}} + \delta \tag{22}$$

$$s_y(m,\bar{\nu}) = \frac{w_{q(m)}}{2\pi\bar{\nu}} = \frac{q(m)}{\Lambda_y\bar{\nu}} + \sigma \tag{23}$$

where $p(m)$ and $q(m)$ describe the one-dimensional ordering of (p, q) onto m. These equations describe the relationship between the periodicity of the object (Λ_x and Λ_y) and the transverse propagation direction. The axial propagation factor $s_z^{(\ell)}$ $(m, \bar{\nu})$ is calculated from a dispersion relation.[21] In contrast to the homogeneous layers, at a given transverse spatial frequency, the field in the transversly inhomogeneous sample consists of contributions of different axial propagation constants [compare $s_z^{(\ell)}$ $(m, \bar{\nu})$ in eq 21 to γ_j in eqs 5, 6, and 17–20].

The field in a homogeneous layer is determined by the vectors $\mathbf{B}^{(\ell)}$ $(m, \bar{\nu})$ and $\hat{\mathbf{B}}^{(\ell)}$ $(m, \bar{\nu})$. Just as $4N_F$ coefficients (β_j and $\hat{\beta}_j$) determine the field in the sample layer, transversality conditions[22] reduce the $6N_F$ elements of $\mathbf{B}^{(\ell)}$ $(m, \bar{\nu})$ and $\hat{\mathbf{B}}^{(\ell)}$ $(m, \bar{\nu})$ to $4N_F$ independent parameters. Thus, for a sample with L layers, $4LN_F$ parameters fully describe the field. As with the case considered in the preceding article, continuity of the transverse electric and magnetic fields can be enforced at the boundaries for each transverse spatial frequency to give $4(L - 1)N_F$ independent constraints. By construction, illumination comes from only one side of the sample so the condition

$$\hat{\mathbf{B}}^{(L)}(m,\bar{\nu}) = \mathbf{0} \tag{24}$$

eliminates another $2N_F$ unknowns. The remaining $2N_F$ parameters are determined by setting the illumination vectors $\mathbf{B}^{(1)}(m, \bar{\nu})$, as described in the first article.[5] The detection of light scattered from the sample is also the same as in the previous article.

For the homogeneous layers considered in the first article, the continuous plane wave spectra[19] could be evaluated numerically by discretizing the transverse propagation cosines s_x and s_y on any grid. In the formulation described here, a natural discrete grid is set by the periodic extension of the object on length scales Λ_x and Λ_y (note that these may be chosen such that the focused light is localized within a single period). This grid may be more coarse than desired, particularly for small values of $\bar{\nu}$. However, the discretization of the incident field in the planewave basis, that is the discretization of (s_x, s_y) in eqs 22 and 23, may be performed for multiple values of (δ, σ). The resulting fields for each value of (δ, σ) may be summed, giving a discretization of (s_x, s_y) on an arbitrarily fine grid, that is at least as fine and commensurate with the descretization dictated by Λ_x and Λ_y.

SIMULATION AND PREDICTION

Numerical simulations are presented here to demonstrate how diffraction and scattering effects in heterogeneous samples are

[21] Equation 5 in ref 5

[22] Equations 6 and 7 in ref 5.

coupled to the sampling geometry, sample morphology, and spectral profile of the sample such that the bulk or so-called “pure phase” spectrum is changed. In the preceding article, it was seen that transmission microspectroscopy is less sensitive to optical distortions than transflection microspectroscopy when a homogeneous layered sample is considered. Further, an overwhelming majority of studies using IR imaging are conducted in the transmission mode.[23] For these reasons, the transmission mode is considered exclusively in the following examples. The extension to transflection is straightforward.

Measurements from two samples are simulated to demonstrate the potential distortions and estimate their magnitude in a first principles manner. In the first example, an object whose response is constant across all wavelengths is considered. Investigation of focused fields in the sample and at the detector illustrates how the spatial structure of the sample affects measurements, independent of the influence of spectral changes. In the second example, full spectral data are simulated for a hypothetical sample of spatially structured toluene, illustrating the increased complexity when spectral variations are added to the sample structure. Effects resulting from the spatial structure of the sample can be seen, and the associated influence on recorded spectra are investigated. In both examples, the effect of an edge on the microspectroscopy data is further investigated. While sensitivity to only the imaginary (absorptive) part of the refractive index is desired, the thickness of the sample and the real part of the refractive index are both seen to affect the data through scattering and diffraction. These effects result in changes in the observed spectral features, including changes in the absorption band profiles and peaks and also changes in the ratios between absorption peaks, which are all quantified.

Frequency-Invariant Sample. In this first example, the sample material considered has no variation in optical response as a function of wavelength. By investigation of the interaction of this sample with focused light of differing wavelengths, some basic behaviors of the microspectroscopy system can be identified. The sample considered is a rectangular slab of absorbing material with index $n = 1.4 + 0.07i$ mounted on a substrate of index 1.45 (i.e., the geometry shown in Figure 1). The slab is 100 μm wide in the x direction and of infinite extent in the y direction, and various thicknesses b in the z direction are considered. The area of interest is taken to be $\Lambda_x = 200$ μm wide in the x direction and infinite in the y direction. The sample is illuminated through the substrate with a y-polarized line-focus. A line-focus is constructed by considering only the $s_y = 0$ line of the aplanatic Cassegrain angular spectrum.[24] A Cassegrain with numerical aperture of 0.5 and a central obscuration aperture of 0.1 is considered. In representing both the object and the field, 200 Fourier series coefficients were retained, i.e., $N_U = N_F = 200$. This level of detail gives sharp edges in the representation of the sample, while increasing the number of Fourier terms did not significantly change the simulation results, indicating that 200 coefficients are sufficient to represent the field. The offsets (δ, σ) were dithered so that there were at least 50 sample points within the numerical aperture of the Cassegrain for all values of $\bar{\nu}$. The angular spectrum from this discretization level leads to a smooth and reasonable focused field.

The line-focus is centered on the absorbing slab in Figure 2. It should be noted that refraction in the substrate has the effect of shifting the nominal focal point.[25] Hence, the sample and substrate have been moved in the axial direction here so that focusing is achieved in the sample plane. This wavelength-dependent (chromatic) shift of the focus has been noted to be a significant problem for dispersive substrates.[25,26] Here it is noted that the substrate also introduces aberration,[25,27] as can be seen by comparing the fields of Figure 2 to fields without a substrate (Figure S1 in the Supporting Information). When the line-focus is positioned between the absorbing slabs, the results of Figure 3 are obtained, while focusing onto the edge of a slab gives the fields shown in Figure 4.

Several comments apply to all three line-focusing cases. Since the sample and illumination have no spatial variation with y and the illuminating light is y-polarized, the field in the sample is also strictly y-polarized. Thus the plots shown are a complete representation of the field. The theory does encompass more general cases, e.g., x-polarized illumination or two-dimensionally focused fields, but the resulting vector fields are more challenging to display. The magnitude of the angular spectrum $B_y^{(3)}(s_x, \bar{\nu})$ is shown in subplots j−l. These spectra can be interpreted as representations of the field strength as a function of direction of propagation. The fine oscillations observed in many of these functions can be attributed to interference between unscattered light and contributions scattered from edges of the slab. Any components of the angular spectrum that lie outside the collection angle of the detection Cassegrain are not collected upon detection. This range is marked by the empty instrument response (i.e., the instrument response with no sample or substrate), in this case $0.1 < |s_x| < 0.5$. Any light diffracted outside the collection range leads to an apparent absorption, as this light is not detected. It should also be noted that any components at $|s_x| > 1$ correspond to waves that are evanescent in free space and do not propagate to the detector. The intensity of light on the detector plane can be calculated from the emerging angular spectra, as described in the previous article.

For illumination focused into the center of the slab, fields within the sample and the transmitted angular spectra are shown in Figure 2. The penetration of the field through the sample is as expected, thicker samples produce more attenuation and longer wavelengths (i.e., lower values of $\bar{\nu}$) are more weakly absorbed. Standing wave effects due to reflection off the top of the sample are also clearly visible. For the thin sample ($b = 2$ μm) it can be seen that there is minimal loss of intensity due to diffraction out of the collection optics, while for thicker samples more light escapes the collection cone. It should be noted that recorded spectra in microspectroscopy are usually of lower signal-to-noise ratio than the bulk recording case. Hence, absorbance of samples is sought to be maximized by adjusting the sample thickness such that the absorbance is maximized in the linear regime of Beer’s law. The typical thickness for most samples is 5−10 μm and feature sizes in many composites and biomedical samples are of

[23] Koenig, J. L.; Wang, S.-Q.; Bhargava, R. *Anal. Chem.* **2001**, *73*, 360A–369A.

[24] Figure 4 in ref 5.

[25] Carr, G. L. *Rev. Sci. Instrum.* **2001**, *72*, 1613–1619.

[26] Wetzel, D. L. *Vib. Spectrosc.* **2002**, *29*, 291–297.

[27] Török, P.; Varga, P.; Laczik, Z.; Booker, G. R. *J. Opt. Soc. Am. A* **1995**, *12*, 325–332.

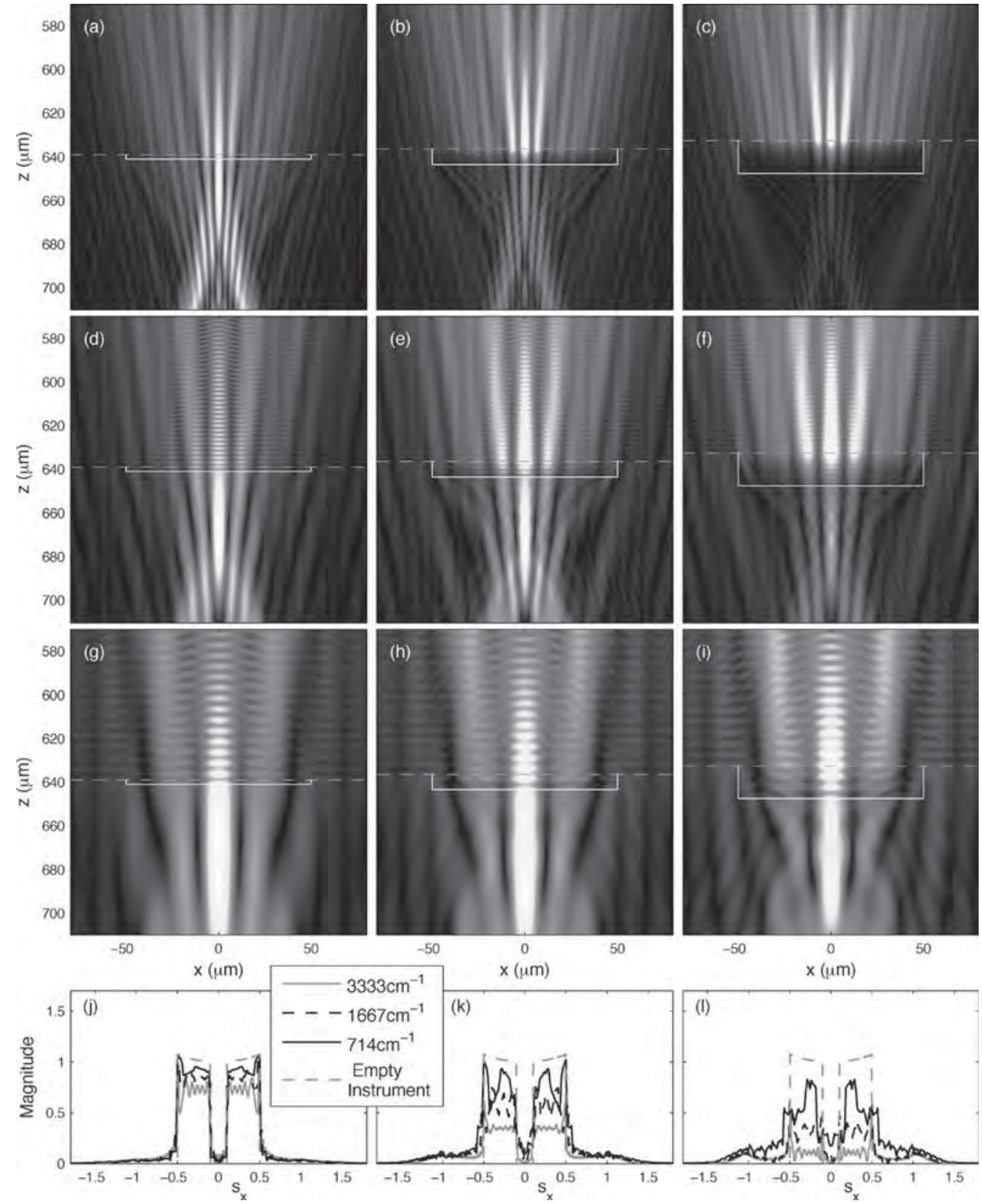


Figure 2. Responses for a line-focused *y*-polarized field incident on the center ($x = 0$) of an absorbing slab. The slab has a complex index 1.4 + 0.07*i* and is mounted on a substrate (the upper region of the plots) of index 1.45 and thickness 2 mm. The field is focused to the $z = 0$ plane in free space. Focusing through the substrate has the effect of moving this focus by about 640 μm, as shown, and also introducing aberration (cf., Figure S1 in the Supporting Information, which considers the same scenario but with the sample suspended in free space). Three sample thicknesses are considered, 2 μm in the left column, 7 μm in the center column, and 15 μm in the right column, that span the usual range in transmission measurements. The *y*-polarized field (the only nonzero field direction) in the sample is shown in parts a−i. The substrate boundary is marked with a dashed line and the slab boundaries with solid lines. Wavelengths of (a−c) 3 μm ($\bar{\nu} = 3333$ cm^{-1}), (d−f) 6 μm ($\bar{\nu} = 1667$ cm^{-1}), and (g−i) 14 μm ($\bar{\nu} = 714$ cm^{-1}) are shown. The magnitude of the angular spectrum after the sample, $B_y^{(3)}(s_x, s_y, \bar{\nu})$ is shown in parts j−l.

a similar order of magnitude. The unfortunate coincidence of order of magnitude for wavelengths, sample features, and optimal path length has an impact on the recorded data for most cases. As this simulation demonstrates, a trade-off between random error and systematic distortion due to optical effects may be avoided in some cases by using thinner samples.

In Figure 3 the same system is considered but with the illuminating light focused between two slabs. There is little light incident on the absorbing material and, apart from a reflection at the substrate boundary, the focused illumination passes through the system largely unperturbed. However, for the thicker samples some scattering effects can be seen in the resulting angular spectra. This illustrates how the optical effects produced by an edge may have a wider region of influence for thicker samples. The implication for a heterogeneous material is that the influence of domains could extend well beyond their obvious morphologic boundaries and proximal regions in a manner that is coupled to the thickness of the sample. While dual aperturing is used in point microspectroscopy to alleviate these effects to some degree, they will be readily apparent in full-field of view imaging.

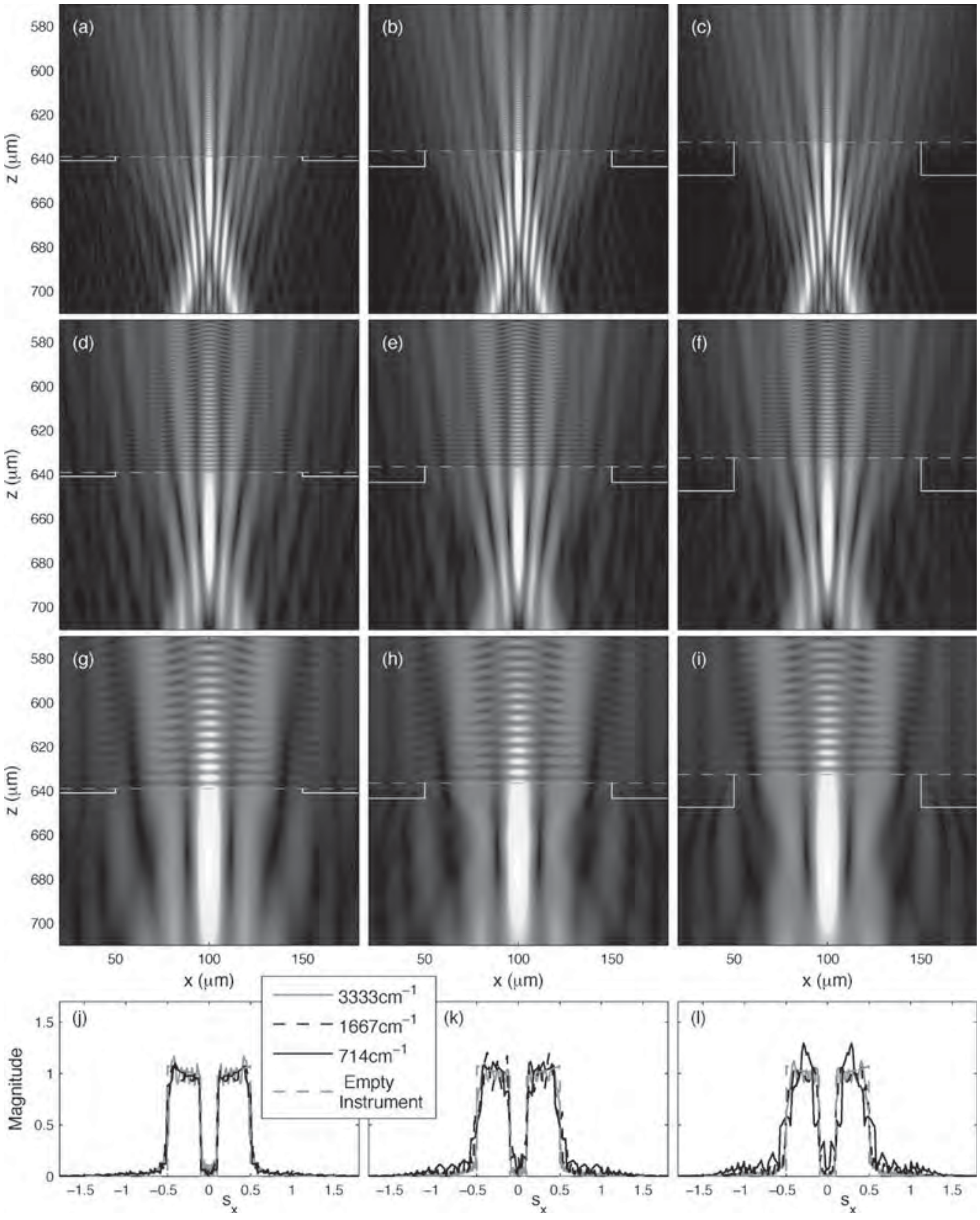


Figure 3. Responses for a line-focused *y*-polarized field incident between two absorbing slabs ($x = 100$ μm) separated by a distance several-fold the wavelength. All other plots and parameters are the same as in Figure 2. A similar scenario, but with the sample suspended in free space, is illustrated in Figure S2 of the Supporting Information.

The illuminating light is focused onto the edge of the sample in Figure 4. In this case, as expected, significant changes to the focused field can be observed. Some of the light is refracted into the absorbing slab and bent out of the collection cone (this can be seen in parts j−l particularly clearly). The resulting sample-induced effects can be seen to trend progressively more prominent with increasing sample thickness. The net effect of an edge is to redistribute spatially the total intensity that would otherwise be incident on the detector. If the distribution is outside of the collection cone, the total intensity reaching the detector is decreased and consequently the apparent absorption is increased. This apparent increase in absorption is only due to optical effects however and depends on the sample morphology. For nonabsorbing spectral regions, the resulting imaging contrast is strong at the edges of domains and is akin to that observed in optical microscopy. The contrast between domains is dictated by their respective refractive indices. While the obvious implication is that an IR microspectrometer may be used in the manner of an optical microscope with properties in the mid-IR region, such a use is not very practical. The primary motivation for working in the mid-IR region is to obtain chemical contrast using absorbance of specific chemical species in spatial domains. Hence, the more important implication is that scattering from nonabsorbing regions of one domain can influence the data recorded in an absorbing spectral region for another domain. In this manner, optical effects complicate data interpretation and make measurements of the spectrum dependent on sample structure.

Animations showing the interactions of the line-focus with the sample are included in the Supporting Information. There is an animation for each combination of wavenumber and sample

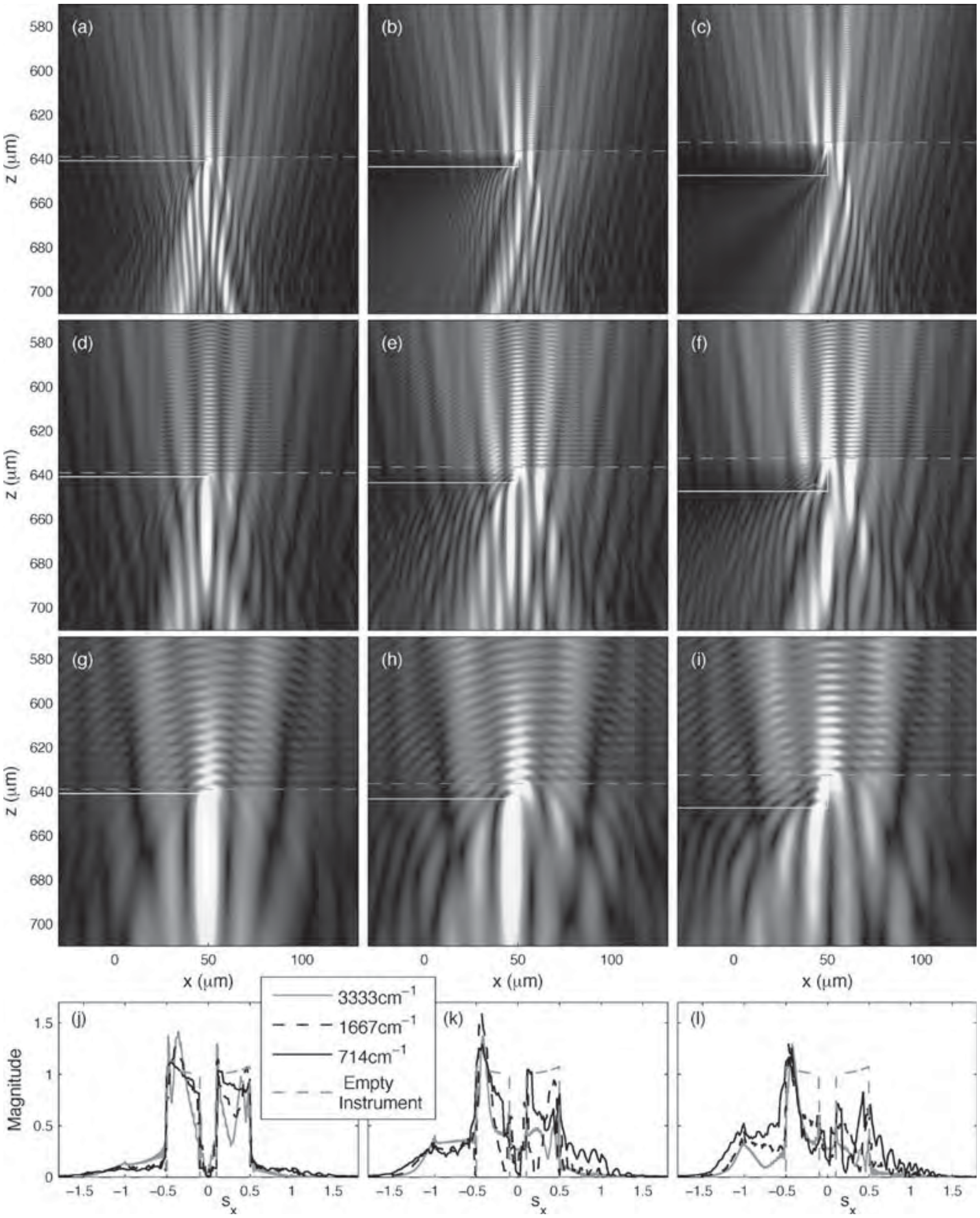


Figure 4. Responses for a line-focused *y*-polarized field incident on the edge ($x = 50$ μm) of an absorbing slab. All other plots and parameters are the same as in Figure 2. A similar scenario, but with the sample suspended in free space, is illustrated in Figure S3 of the Supporting Information.

thickness seen in Figure 2, and a second animation for each combination but with the sample suspended in free space rather than on a substrate.

When the diffracted components are collected, the distribution of light intensity in the detector plane is also affected, meaning that contributions from the edge effects can produce artifacts in pixels besides the ones associated with the edge position. To understand the practical effects of spatial redistribution, a fully two-dimensional focusing solution is needed. Hence, a full focusing aperture, rather than a line-focus, is considered for the remainder of this article. In all cases, circular apertures (as shown in the previous article[24]) are represented on a discrete Cartesian grid, as consistent with the analysis presented in the previous section. The effects of light redistribution are illustrated in Figure 5.

The object from Figure 2 is considered in Figure 5 but represented with $N_U = 40$ coefficients. The angular spectrum of illumination is discretized so that for any wavenumber $\bar{\nu}$ the s_x diameter across the aperture is at least 20 pixels and the s_y diameter is 20 pixels. The field emerging from the sample is represented using an angular spectrum discretized with the same pixel spacing and with 20 pixels in the s_y dimension and at least 60 pixels in the s_x dimension. The discretization described here is more coarse than that used in the previous calculations of the fields in the sample. This is because the predicted detection data are less sensitive to fine features of the field (e.g., evanescent waves) so that the desired prediction ceases to change with the discretization at a more coarse level. The outer and inner Cassegrain numerical apertures are again

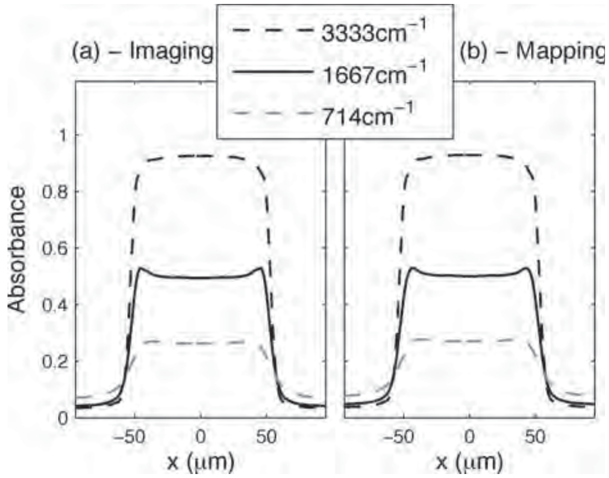


Figure 5. Absorbance profiles of the 7 μm slab of Figure 2, at different wavenumbers and for both (a) imaging and (b) point mapping modalities.

0.5 and 0.1, respectively, which means that the discrete representation of the scattered fields extends well into the evanescent region.

Two modalities were simulated. First, the focus of an unpolarized illuminating field was translated in small increments and the emerging angular spectrum calculated. By calculation of the total power throughput,[28] a point mapping system was simulated, where a large area detector was used and the Cassegrain edges set the limiting apertures in the optical path. That is, the sample was illuminated at a single spot and the transmitted light captured using a single IR detector. Second, widefield illumination with array detection was simulated. The transmitted angular spectrum can be used to calculate the intensity on a detector plane,[29] and for each focal position these intensities can be summed. The fill factor of the detector is not explicitly considered here but the consequences of a nonunity fill factor can be included in the model and are not expected to produce significant qualtitative changes. These two approaches, point mapping and imaging, are both employed in contemporary microspectroscopy, and simulations in Figure 5 for both modalities demonstrate similar results. However, it is instructive to notice how the measured profile of the slab depends on wavenumber. For example, both the gradient of the absorbance at the slab edge and the overshoot at the edge vary with wavenumber. While the wavenumber dependence of the achievable spatial resolution is known,[30] spectral measurements also change with wavelength due to optical effects (such as diffraction and refraction) and, additionally, with sample structure (e.g., thickness). A description of spectral distortions and their effect on spatial specificity (and, in turn, the resolution attainable) is lacking. The model sample of constant $k(\bar{\nu})$ considered here exhibits differing profiles due to wavelength dependent phenomena, emphasizing this relationship between recorded spectra and the apparent morphology of the sample.

Frequency-Variant Samples. To see how optical phenomena influence a measured spectrum, the simulation parameters described for Figure 5 were modified by replacing the constant index of the slab with the complex refractive index of toluene[31] and by replacing the constant index of the substrate by the index of barium fluoride.[32] A background measurement was calculated by applying standard transmission coefficients to model the transmission of light through the air-to-barium-fluoride boundary at the first substrate surface and the barium-fluoride-to-air boundary at the second substrate surface. In the presence of the absorbing toluene slab, both point mapping and imaging profiles were calculated using the methods described above.

Spectra from the imaging modality are shown in Figure 6. In these calculations it was assumed that the pixel size was 5 μm at the sample plane. Spectra are plotted for the center of the slab and for measurements in the vicinity of the edge. It can be seen that light scattered outside of the collection cone produces a nonzero baseline in the measured spectra, as is commonly observed. A smooth baseline function is often fitted to these spectra and subtracted out before spectral metrics are calculated. Here, local linear baselines are fitted to the spectra, as is common practice in spectral preprocessing, and peak position and height metrics calculated (as illustrated in Figure 6). The resulting peak positions are given in Table 1, and the resulting normalized peak heights are given in Table 2. In both cases an ideal value has been calculated by determining the true absorbance profile from the imaginary refractive index.[33]

It can be seen that the observed spectral metrics depend on the position at which the spectra are measured. Optical effects distort the spectra by coupling the real part of the refractive index and the sample structure into the data. While baselining has removed some of the gross optical effects, the metrics are not independent of morphology. It should be noted that correction algorithms other than baseline subtraction have been proposed, e.g., taking derivatives of the spectra or more advanced procedures.[34,35] However, these procedures are typically ad hoc or do not fully account for physical phenomena such as the coupling of the dispersive line-shape (the real index) into the observed spectra and the influence of the sample morphology on the collected data. Hence they cannot capture the physics of the true distortions and may provide unjustified confidence compared to uncorrected data.

The point mapping modality was also simulated, and the measured spectra are shown in Figure 7. While there are differences, the gross behavior can be seen to be similar to that observed in the imaging modality. In this example, the observed peak positions (Table 3) are the same as for the mapping case, while the peak ratio (Table 4) metrics differ but exhibit a similar amount of variability as the imaging case. The baseline characteristics differ between the imaging and mapping modalities. This is to be expected as scattering distorts the point spread function of the light to spatially redistribute light intensity incident on the detector—in imaging mode this means that light scattered from an edge can effect neighboring pixels, while for mapping this type of crosstalk does not occur.

[28] Equation 32 in ref 5.
[29] Equations 30 and 31 in ref 5.
[30] Lasch, P.; Naumann, D. *Biochim. Biophys. Acta* **2006**, *1758*, 814–829.
[31] Figure 6 in ref 5.
[32] Malitson, I. H. *J. Opt. Soc. Am.* **1964**, *54*, 628–632.
[33] Equation 35 in ref 5.
[34] Kohler, A.; Kirschner, C.; Oust, A.; Martens, H. *Appl. Spectrosc.* **2005**, *59*, 707–716.
[35] Thennadil, S. N.; Martens, H.; Kohler, A. *Appl. Spectrosc.* **2005**, *60*, 315–321.

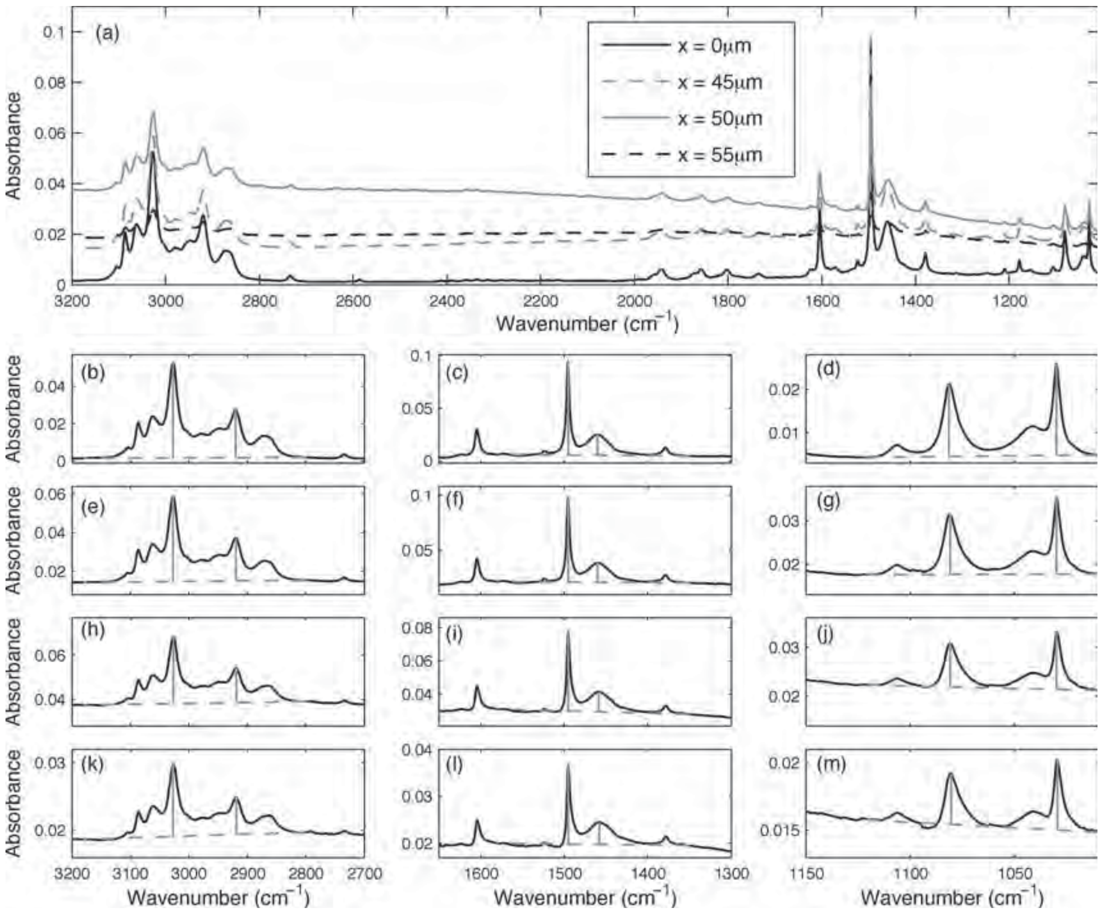


Figure 6. Imaging spectra from a 7 μm thick toluene slab on 2 mm of barium fluoride. The absorbance is normalized by the slab thickness. Spectra are shown from the center of the slab ($x_0 = 0$) and in the vicinity of the edge ($x = 50$ μm). The full spectra (a), and details for $x = 0$ (b−d), $x = 45$ μm (e−g), $x = 50$ μm (h−j), and $x = 55$ μm (k−m) are shown. A baseline is illustrated by a dashed line in the detail plots, and peak heights and positions are calculated as illustrated by the solid vertical lines. The calculated metrics are given in Tables 1 and 2.

Table 1. Peak Positions for the Imaging Data to the Nearest 0.5 cm^{-1}

	peak 1	peak 2	peak 3	peak 4	peak 5	peak 6
ideal	3027.0	2920.0	1495.5	1460.5	1030.0	1081.5
$x_0 = 0$	3027.0	2920.0	1495.5	1460.0	1030.0	1081.5
$x_0 = 45$ μm	3027.0	2919.5	1495.5	1460.0	1030.0	1081.0
$x_0 = 50$ μm	3026.5	2919.0	1495.5	1459.0	1030.0	1081.0
$x_0 = 55$ μm	3027.0	2919.0	1495.0	1458.0	1029.5	1080.5

Table 2. Normalized Peak Heights for the Imaging Data

	peak 1	peak 2	peak 3	peak 4	peak 5	peak 6
ideal	1.00	0.514	1.79	0.436	0.444	0.350
$x_0 = 0$	1.00	0.505	1.76	0.388	0.416	0.327
$x_0 = 45$ μm	1.00	0.505	1.74	0.400	0.401	0.314
$x_0 = 50$ μm	1.00	0.517	1.57	0.401	0.388	0.296
$x_0 = 55$ μm	1.00	0.517	1.58	0.447	0.486	0.352

Note that the severity of the metric distortion depends on the sample morphology and boundaries. For example, Figures S4 and S5 in the Supporting Information show results for a sample thickness of 2 μm rather than 7 μm. It can be seen that optical distortions, such as the nonzero baseline, are less severe for the 2 μm thick sample. As noted earlier, thinner samples can, in general, be expected to be less susceptible to distortions due to optical phenomena than comparable thicker samples. Spectral metrics are also affected to a lesser extent as can be seen by comparing the metric tables for the 2 μm thick sample (Tables S2−S5 in the Supporting Information) to the metric tables for the 7 μm thick samples above. For example, in the latter, a maximum peak shift of 2.5 cm^{-1} is observed, while for a 2 μm thick sample, the maximum peak shift is 1 cm^{-1}.

The dependence of spectral distortions on sample parameters is important from two perspectives. First, the effect of geometry becomes difficult to quantify in simple terms. Hence, a measure of the systematic deviations in the spectrum must be individually calculated for specific samples. This is especially important for studies that are interested in subtle chemical changes at edges (often several wavenumber shifts) or in an algorithm-based search. While careful simulations are prescribed for sensitive chemical analyses, the strategy in database searching may be to use a coarse spectral resolution. Second, in automated analysis algorithms such as those for tissue histopathology,[36] sample thickness becomes an important parameter whose impact must be appreciated. One approach may be to carefully control sample thickness such that deviations are consistent and can be eliminated from use in classification algorithms by choice of appropriate metrics. A second approach is to use a large number of samples with a thickness variation arising from the natural variation of the

[36] Fernandez, D. C.; Bhargava, R.; Hewitt, S. M.; Levin, I. W. *Nat. Biotechnol.* **2005**, *23*, 469–474.

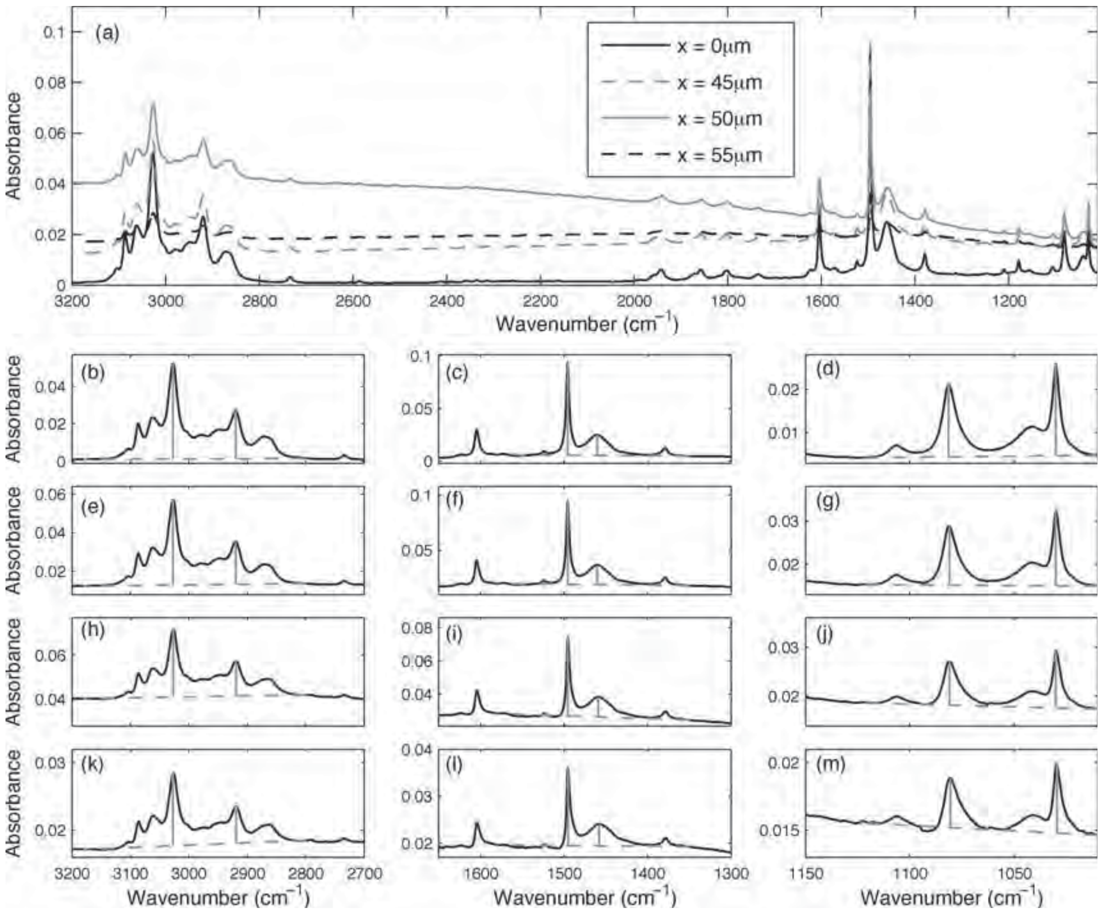


Figure 7. Point mapping spectra from a 7 μm thick toluene slab on 2 mm of barium fluoride. The absorbance is normalized by the slab thickness. Spectra are shown from the center of the slab ($x = 0$) and in the vicinity of the edge ($x = 50$ μm). The full spectra (a) and details for $x = 0$ (b−d), $x = 45$ μm (e−g), $x = 50$ μm (h−j), and $x = 55$ μm (k−m) are shown. A baseline is illustrated by dashed lines, and peak heights and positions are calculated as illustrated by the solid vertical lines. The calculated metrics are given in Tables 3 and 4.

Table 3. Peak Positions for the Point Mapping Data to the Nearest 0.5 cm^{-1}

	peak 1	peak 2	peak 3	peak 4	peak 5	peak 6
ideal	3027.0	2920.0	1495.5	1460.5	1030.0	1081.5
$x_0 = 0$	3027.0	2920.0	1495.5	1460.0	1030.0	1081.5
$x_0 = 45$ μm	3027.0	2919.5	1495.5	1460.0	1030.0	1081.0
$x_0 = 50$ μm	3026.5	2919.0	1495.5	1459.0	1030.0	1081.0
$x_0 = 55$ μm	3027.0	2919.0	1495.0	1458.0	1029.5	1080.5

Table 4. Normalized Peak Heights for the Point Mapping Data

	peak 1	peak 2	peak 3	peak 4	peak 5	peak 6
ideal	1.00	0.514	1.79	0.436	0.444	0.350
$x_0 = 0$	1.00	0.502	1.75	0.386	0.415	0.326
$x_0 = 45$ μm	1.00	0.504	1.73	0.403	0.397	0.309
$x_0 = 50$ μm	1.00	0.520	1.56	0.395	0.382	0.292
$x_0 = 55$ μm	1.00	0.509	1.55	0.435	0.475	0.342

protocol. Any developed classification algorithm then will be insensitive to optics-induced distortions within the range of thicknesses used in the development of the protocol.

Experimental Comparison. To test the predictive power of the model presented here, it is useful to compare experimental data with simulations for a comparable sample and imaging system. The sample data were recorded on a Varian Stingray system using a mid-IR interferometer. The microscope of the instrument is equipped with a narrowband, liquid nitrogen cooled mercury−cadmium−telluride (MCT) detector, as well as a 128 × 128 pixel, liquid nitrogen-cooled focal plane array MCT detector. Data are recorded at an undersampling ratio of 2 referenced to the He−Ne laser, zero-filled by a factor of 2, and Fourier transformed using Happ−Genzel apodization. The nominal spectral resolution was 2 cm^{-1}. The ratios of two similarly collected image sets (one without a sample to serve as a background and one with a sample) are taken pixel by pixel to obtain absorbance image datasets. A common photoresist material, SU-8 2000.5 (MicroChem Corp., Newton, MA), was spin coated to an approximate thickness of 10 μm on a 25 mm diameter barium fluoride (BaF_2) disk and pattern cured by UV exposure using a standard USAF 1951 target (Edmond Optics, Barrington, NJ). The entire sample was baked at 95 °C and developed as per standard protocols for postcuring. A postbake at 150 °C for 5 min was performed to ensure complete polymerization and long-term stability.

An image of the transmittance, at $\bar{\nu} = 2903$ cm^{-1}, for a region of the target is shown in Figure 8. The data measured along the dashed line will be examined; in particular, the spatial−spectral response across the edge of a bar structure is of interest. The absorbance profile along the dotted line shown in Figure 8 is

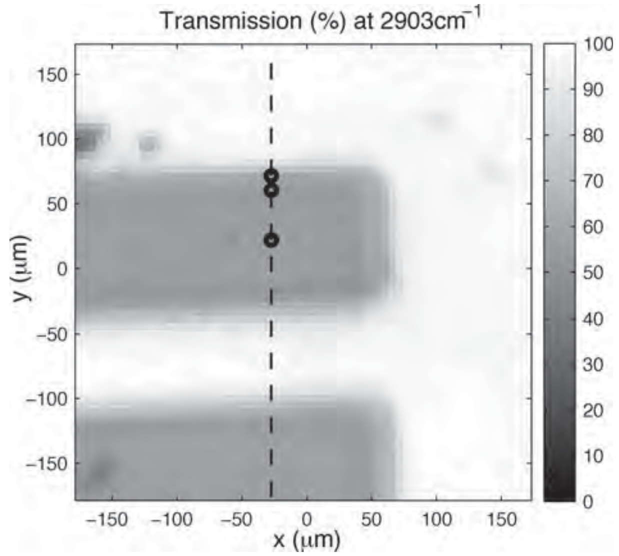


Figure 8. Transmission image of a SU-8 bar target on barium fluoride at 2903 cm^{-1}. In subsequent figures, profiles will be displayed from along the dashed line, and spectra will be plotted for the points marked with a circle.

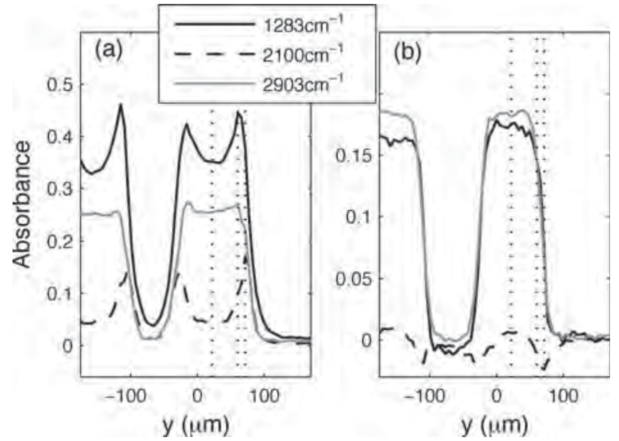


Figure 9. Absorbance profiles, before (a) and after (b) baseline correction, across the bar target for three different wavenumbers. At $\bar{\nu} = 1283$ cm^{-1} and $\bar{\nu} = 2903$ cm^{-1}, the SU-8 polymer is absorbing. At $\bar{\nu} = 2100$ cm^{-1}, the polymer is nonabsorbing but scattering effects produce apparent absorption at the edges.

plotted for three different wavenumbers in Figure 9a. Apparent artifacts, e.g., overshoot in the absorbance at the sample edges, can be seen to vary with wavenumber. These distortions arise from both baseline offset due to redistribution of intensity by the sample and changes in the apparent peak shape. Other wavenumber-dependent effects are also visible, e.g., the change in spatial resolution as a function of wavenumber is manifest in the differing gradients of the absorbance profiles at the edge.

Subtracting a slowly varying baseline is a common method to compensate for the consequences of optical effects on spectra. In Figure 9b, the edge profiles are replotted after a linear baseline has been subtracted from the spectra. For each of the spectra, the baseline was found by linear interpolation between minima of the SU-8 response, specifically between the absorbance values at 910, 1423, 1551, 1827, 2696, 2783, 3111, 3736, and 3931 cm^{-1}. It can be seen that the baselining procedure qualitatively improves the edge profiles, at least in absorbing regions of the spectrum. Such subjective baselining can lead to seemingly reasonable results, especially when scattering is high and absorbance is low. For automated analyses, which are required due to the large number of pixels (spectra) making manual correction impossible, simple corrections may lead to errors. For example, at 2100 cm^{-1}, the baselining procedure has resulted in some nonphysical negative values of absorbance. Another potential concern is the discrepancy in absorbance between the two bar targets. For the bar centered around $y = 20$ μm, the absorbance values at 1283 and 2903 cm^{-1} are approximately equal, while the neighboring bar exhibits a greater difference, despite being made of the same material and being subject to the same processing history.

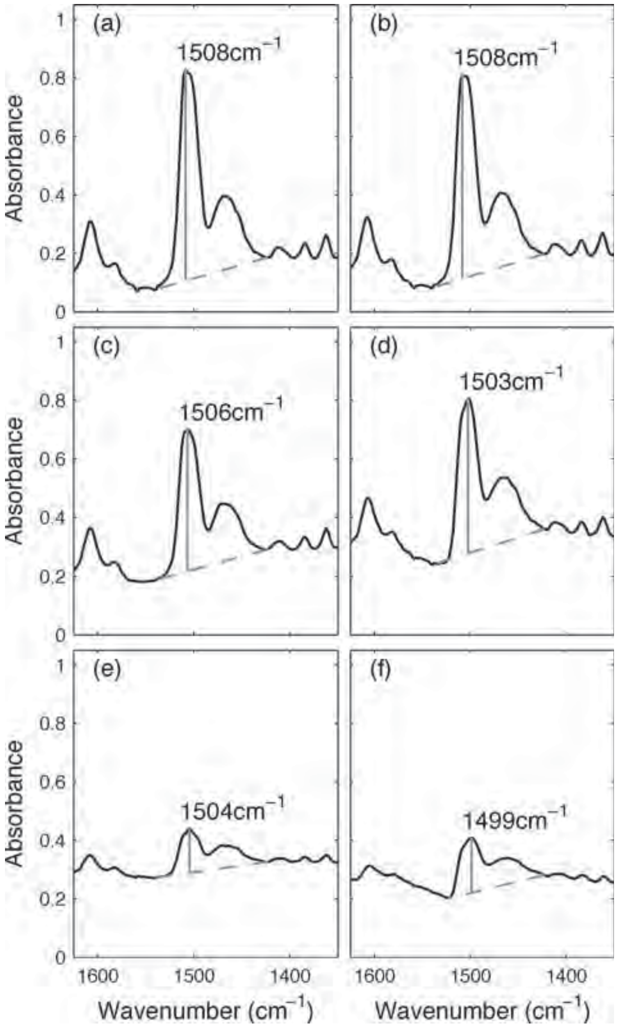


Figure 10. Experimental absorbance spectra taken at $y = 22$ μm (a), $y = 60.5$ μm (c), and $y = 71.5$ μm (e) (i.e., the points marked with a circle in Figure 8) and simulated spectra taken from the polymer 50 (b) and 6 μm (d) from the edge and from off the polymer 5 μm from the edge (f). The peak location, after the illustrated baseline correction, is displayed on the plots.

Quantitative examination of the collected data reveals spectral distortions of the type predicted earlier in the article. An illustrative absorption peak is centered around $\bar{\nu} = 1508$ cm^{-1}. Experimental measurements of this peak are shown for various sample locations in the left column of Figure 10. Data collection from this peak can be simulated by first estimating the physical properties of the sample. By comparison of the absorbance measured at $y = 22$ μm in relation to the imaginary part of the refractive index of SU-8 calculated in the previous article, the thickness of the SU-8 was estimated to be approximately 7 μm. The imaginary part of the refractive index was then estimated from the absorbance[33] (again from the measurement at $y = 22$ μm). Kramers−Kronig[37] analysis was used to calculate the real part of the refractive index, thus completing the description of the object. Note that the SU-8 refractive index calculated in the previous article was not employed, as differences in sample preparation were found to have introduced small but significant differences in the optical properties of the polymer.

The sample edge profile and the instrument were modeled in the same manner used to generate Figure 6, except that the inner and outer numerical apertures of the Cassegrain were taken to be 0.26 and 0.4, respectively. These values are consistent with those used for the same instrument in the previous article. The experimental and predicted spectral profiles of parts a and b of Figure 10 agree well. This is to be expected as the SU-8 refractive index used in the simulations was calculated from Figure 10a, and this region of the sample is a relatively simple layered structure.

In the vicinity of the polymer edge, the peak position in the experimental data can be seen to shift. Since the target structure is made of a single material, this shift can most likely be attributed to optical effects. Nonuniform curing occurring at the sample edges can be ruled out due to extensive postreaction thermal cure. The simulations also predict a peak shift toward lower wavenumber; however, this shift is greater in the predictions than it is in the measurements. There are several possible causes for this overestimation. The characterization of the sample relied on a chain of estimation procedures, the real index was estimated from the imaginary index which was in turn dependent on the assumed sample thickness, with the possibility of propagating errors. The correct prediction of bulk spectra, however, suggests that this error is small. The sample geometry may also lead to errors in prediction. In simulation, the edge is represented by a steep gradient between two perfectly flat surfaces. In reality, the sample edges can be expected to have some finite and unknown slope, and the horizontal surfaces in the bar targets may not be perfectly flat. The broad agreement between the experimental and simulated results of Figure 10 indicate that the model developed has significant predictive power and allows an understanding of the causes and effects of optical artifacts.

CONCLUDING REMARKS

This article presents the first attempt at applying rigorous optical theory to heterogeneous samples in IR microspectroscopy. It is shown that lateral structure in thin samples leads to significant effects on the recorded spectral data arising from a coupling between wavelength, sample geometry, optical properties within the sample, presence of interfaces, and the optical setup. With the use of progressively sophisticated simulations, the effect of each of these factors was demonstrated in a quantitative manner. It was shown that the redistribution at the detector place of the intensity incident upon the sample can be quantitatively modeled and verified with experiments. The implications for the practice of spectroscopy are that the spatial and spectral variation of the real and imaginary parts of the index of the sample cannot be decoupled from FT-IR imaging data, as is currently practiced. It is emphasized that recording the true data will require the development of both new instruments that can provide additional data to extract true spectral properties from the data, as well as numerical methods to assist in the same. The theoretical framework presented here should serve as a useful guide to estimate the true structure and quantify distortions in present instruments as well as a platform for future development.

ACKNOWLEDGMENT

This work was supported in part by the Grainger Foundation for Emerging Technologies, the Department of Defense Prostate Cancer Research Program, and Susan G. Komen for the Cure. The authors would also like to thank Dr. Gokulakrishnan Srinivasan for fabricating the SU-8 sample and Dr. Michael J. Walsh for help in acquiring data.

SUPPORTING INFORMATION AVAILABLE

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review September 15, 2009. Accepted December 31, 2009.

AC902068E

[37] Kuzmenko, A. B. *Rev. Sci. Instrum.* **2005**, *76*, 083108.

Accurate histopathology from low signal-to-noise ratio spectroscopic imaging data

Rohith K. Reddy and Rohit Bhargava*

Received 25th May 2010, Accepted 22nd July 2010
DOI: 10.1039/c0an00350f

Fourier Transform Infrared (FT-IR) spectroscopic imaging is emerging as an automated alternative to human examination in studying development and disease in tissue. The technology's speed and accuracy, however, are limited by the trade-off with signal-to-noise ratio (SNR). Signal processing approaches to reduce noise have been suggested but often involve manual decisions, compromising the automation benefits of using spectroscopic imaging for tissue analysis. In this manuscript, we describe an approach that utilizes the spatial information in the data set to select parameters for noise reduction without human input. Specifically, we expand on the Minimum Noise Fraction (MNF) approach in which data are forward transformed, eigenimages that correspond mostly to signal selected and used in inverse transformation. Our unsupervised eigenimage selection method consists of matching spatial features in eigenimages with a low-noise gold standard derived from the data. An order of magnitude reduction in noise is demonstrated using this approach. We apply the approach to automating breast tissue histology, in which accuracy in classification of tissue into different cell types is shown to strongly depend on the SNR of data. A high classification accuracy was recovered with acquired data that was ~10-fold lower SNR. The results imply that a reduction of almost two orders of magnitude in acquisition time is routinely possible for automated tissue classifications by using post-acquisition noise reduction.

1. Introduction

Fourier Transform Infrared (FT-IR) spectroscopic imaging[1] with array detectors provides large data sets but often requires large times for acquisition of high signal to noise ratio (SNR) data. Following conventional trading rules in IR spectroscopy,[2] hence, the signal is recorded multiple times and added to increase the signal to noise ratio (SNR) of the data. In imaging, other approaches have also been suggested due to the complex nature of the acquisition process.[3,4,5,6] Fundamentally, these methods unavoidably traded the SNR reduction against an increase in acquisition time. Another approach may be to improve hardware but is expensive and impractical for most users. A final and very successful approach has been to trade off the spatial coverage per scan using sensitive linear array detectors, obviously limiting the spatial coverage rate. For a finite data acquisition time, other schemes to extract low noise information are available[7] but these methods neglect the image as a whole and result in loss of image fidelity. As a consequence, FT-IR imaging data acquisition is limited in applications that require fast imaging at high fidelity.

Using computation to enhance instrument performance is becoming an attractive option with the rapid development of powerful computers and increased storage capacities. A procedure based on the Minimum Noise Fraction (MNF) transform,[8] for example, was adopted from the satellite, airborne and other imaging communities[9] for IR spectroscopic imaging.[10,11] Similarly, ideas in data compression and with the potential for attendant noise reduction are being proposed by other groups.[12,13] In this milieu, a general approach to noise reduction is to use an Eigenvalue decomposition of the data using a forward transform, for example, a principal components analysis (PCA). After selecting eigenimages with sufficient SNR, the selected data are inverse transformed to yield the entire dataset with lower noise content. This approach was used in FT-IR imaging, for example,[14] to examine phase compositions by enhancing contrast between different regions. PCA reorders data in decreasing order of variance. Similarly, techniques can be used to order eigenimages in decreasing order of SNR, which is the aforementioned MNF transform. A modified version[15] of this transform was shown to improve image fidelity and achieve better noise reduction than PCA, for example.

Mathematical transform techniques for noise reduction generally utilize the property that noise is uncorrelated whereas spectra (signals) have a higher degree of correlation. In the transform domain, hence, the signal becomes largely confined to a few eigenvalues whereas the noise is spread across all. Noise reduction can be achieved by retaining eigenvalue images that corresponding to high signal content and computing the inverse transform. It is the relative proportion of the signal and noise which forms a criterion for inclusion of specific eigenimages in the inverse transform. Inclusion of too many will not allow for significant noise rejection, while inclusion of too few would result in loss of fine spectral features. Hence, identifying eigenvalues corresponding to high signal content is an important step in the noise reduction process.[16,17] Most methods[16,18,19] choose the first *m* eigenimages. The value of *m* may be chosen by considering the decay of the information content (eigenvalues). The assumption

Department of Bioengineering and Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, USA. E-mail: rxb@illinois.edu; Fax: +1 (217) 265 0256; Tel: +1 (217) 265 6596

that the first m eigenimages should be chosen, however, is questionable. The MNF approach, for example, was specifically developed to overcome the observation that the first m eigenimages in PCA were not always optimal and proposed, instead, a noise-based ordering. Other methods[17,20] can be computationally expensive or do not utilize some of the features of the data. All methods, hence, place the ordering burden on the decomposition algorithm and do not directly utilize features of the data or the unique features of the image acquisition process. The method proposes here addresses this gap in utilizing data effectively by using the structure within the data to select features.

Another general criticism of present methods is that they do not explicitly account for the correlated spatial and spectral information in the data. The variance in data may arise from measurement noise, sensor characteristics or due to scattering effects from the sample. For example, the MNF approach can be shown to rigorously order images in decreasing order of random noise. Implicitly, the signal in the re-ordering of MNF eigenimages is assumed to arise from features in the image but could come from those other than the sample of interest. We present such a case in Fig. 1, which shows the 4th, 8th, 12th and 19th eigenimages for FT-IR imaging data from a breast tissue sample acquired following procedures previously reported.[21] The 4th eigenimage shows interesting tissue morphological *i.e.* structural features. Although the 8th eigenimage has higher SNR compared to the 12th or 19th, the 12th and 19th eigenimages seemingly contain more features of interest. Obviously, here one would include the 12th and 19th but not the 8th image in a noise reduction scheme. The 8th eigenimage likely arises from water vapor differences, as can be seen by examining the spectra acquired in this data set using a small linear array that is raster scanned horizontally from bottom to top, and not from the sample itself. Hence, for spectroscopic imaging data sets, it may be more instructive to employ a method of selecting eigenimages that accounts for both spectral and spatial correlations.

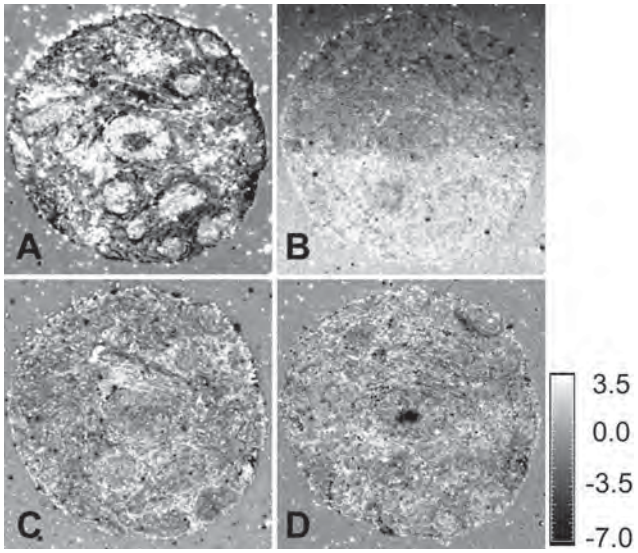


Fig. 1 (A) 4th MNF Factor (Tissue features are apparent) (B) 8th MNF factor (C) 12th MNF factor (D) 19th MNF factor. The 8th factor has less apparent structural features than others and is dominated by measurement artifacts.

There is no universal algorithm to optimally include both spatial structure pertinent to the sample and spectral characteristics in selecting appropriate eigenimages. Hence, the identification of eigenimages to include in the data inversion process is invariably a manual task. This requirement makes the automation advantage limited and is a key impediment to automated and routine application of noise rejection methods. First, though the effect is likely to be small, manually selected eigenimages will likely vary from practitioner to practitioner and may lead to variance in scientific conclusions or confidence in results. Second, the need to examine every eigenvalue image (or, at least, a large set of images) is time-consuming. The decision to exclude or include images with questionable content requires significant time and some other guidance, *e.g.* a complementary optical microscopy image. While such data are often available, they are presently not used in noise rejection. In this manuscript, we propose a method to automatically determine eigenimages to use in an inverse transform for effective noise rejection by enabling the use of additional information to recognize important features in the data. The proposed algorithm selects eigenimages based on structural features in a quantitative manner by utilizing both the correlation between spectra as well as the spatial information in the image. We test the automated noise rejection algorithm by comparing information about tissue structure extracted from data before and after noise rejection. Last, the improvements in SNR are quantified and discussed in terms of potential data acquisition strategies.

2. Methods

2.1. Mathematical background to the proposed method

The MNF transform was introduced by Green *et al.*[8] to order multispectral data in terms of image quality and we briefly describe the background to our approach next. Consider a three-dimensional (3-D) dataset $X_k(\vec{t})$ where $\vec{t} = (i,j)$ represents spatial data coordinates and k denotes the spectral element index. If the number of spectral elements in the data are M, then $X(\vec{t}) = [X_1(\vec{t}), X_2(\vec{t}), X_3(\vec{t}) \ldots X_M(\vec{t})]^T$ and the true spectral value, S and additive noise, N, are related as

$$X(\vec{t}) = S(\vec{t}) + N(\vec{t}) \tag{1}$$

Consequently, the covariances are related through

$$Cov(X) = Cov(S) + Cov(N). \tag{2}$$

Next, the noise fraction for the k^{th} spectral element is defined in terms of the variance of the noise

$$F_k = Var(N_k)/Var(X_k) \tag{3}$$

which is the ratio of noise variance to the total variance of that spectral element. The MNF transform is a linear combination of bands

$$Y_k(\vec{t}) = \sum_{m=1}^{M} \alpha_m^k X_m(\vec{t}) \tag{4}$$

such that the noise fraction F_k is minimum for $Y_k(\vec{t})$ among all linear transformations orthogonal to $Y_j(\vec{t})$, $j = 1, 2, \ldots k$. The vectors $\alpha^k = [\alpha_1^k, \alpha_2^k, \alpha_3^k \ldots \alpha_M^k]^T$are the left hand eigenvectors of $\Sigma_N \Sigma^{-1}{}_X$. Also the eigenvalue corresponding to α^k is equal to the noise fraction of Y_k, *i.e.*

$$\lambda_k = F_k \tag{5}$$

The definition of MNF would imply that $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_M$. Since λ_k corresponds to the noise fraction, MNF re-orders spectral elements in terms of increasing F_k or equivalently, in terms of decreasing SNR. The same set of eigenvectors is obtained from maximizing SNR or the noise fraction. However, the approach that maximizes SNR would result in higher eigenvalues corresponding to higher SNR and the MNF transform would result in decreasing order of SNR corresponds to decreasing order of eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_M$. Our implementation uses this approach to compute MNF transforms. It is useful to note that since the MNF transform depends on signal to noise *ratio* it is invariant under scale changes to any band (unlike principal components).

2.2. Proposed algorithm based on MNF-transform for noise reduction

The MNF transform is first computed following the method above. In heterogeneous materials and tissue, we note that the eigenimages also have structure corresponding to the true structure of the material. The contrast and precise values of any spectral and eigenimage, of course, cannot be equated but both types of images have distinct spatial domains. Domains are defined by their edges and this property forms the basis of our eigenimage selection scheme. Our proposed method relies on leveraging the spatial structure in spectroscopic imaging data with structural details in eigenimages *via* comparisons of domain edge profiles. Domains in breast tissue, for example, include boundaries of the sample, ducts and transitions between different structural units. Several methods for edge detection[22] based on different filters and different thresholding schemes have been proposed and studied. Canny's method,[23] in particular, is widely used and was found to be effective for our application. We evaluated two other edge detection methods (Sobel, Roberts) but found the Canny method better suited to the relative domain and pixel sizes likely because Canny's method has been shown to be optimal with respect to detection, localization and response.[23] The result of edge detection is a binary image that is termed an 'edge map'. A typical absorbance image and edge map is shown in Fig. 2. While it is indeed possible to determine edges for any image in general, confounding effects may arise when the domain sizes are similar to pixels sizes and the images are noisy. Hence, an intermediate step may be to use a median filter. The choice of size of the median filter will be a compromise between the size of structural features and pixel sizes. Using a large median filter would be effective in removing pixel-to-pixel variations but could also result in loss of features, especially those that are smaller than the size of the median filter. Median filters of sizes between 7 × 7 and 13 × 13 were found to be most effective for the samples considered here in that the results were very consistent regardless of the choice of filter. Hence, we elected to routinely use a 9 × 9 pixel filter which is used on the data in Fig. 2 prior to obtaining the edge map.

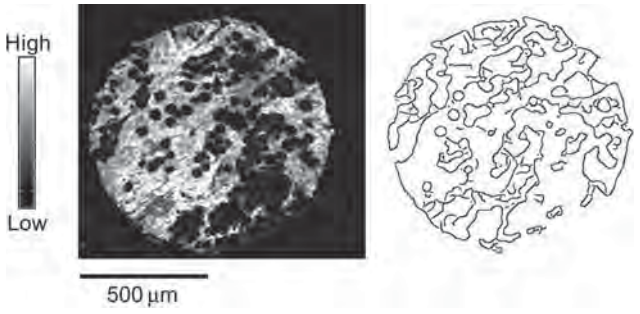


Fig. 2 A typical image (left) obtained by plotting the absorbance of the sample (here at 3400 cm^{-1}) and the corresponding edge map (right) obtained after median filtering and using Canny's method.

Once edge maps for the eigenimages are obtained, we seek to compare them to an 'ideal' image (**I**). It is desirable that this image contains all structural details of interest as well as be of high SNR. The edge map of **I** and edge maps of eigenimages can then be compared and eigenimages that are sufficiently representative of the sample structure may be included in the inversion. The proposed method does not depend on any specific method to generate edge maps and diverse contrast mechanisms can be used to construct the edge map of the ideal image. Several possibilities are discussed next. The first MNF eigenimage corresponds to the highest SNR and likely contains the greatest sample detail; hence, it could be used as an ideal image. Another avenue may be to choose an image based on the molecular characteristics of the sample if prior knowledge about the sample is available. For example, spectral characteristics of tissues for a given organ is often well-constrained in the spectral regions between ∼ 950 cm^{-1} (lower detector cut-off in some of the experiments here) to ∼1800 cm^{-1} (mainly bending and rocking vibrational modes of molecules) and from ∼2765 cm^{-1} to ∼3750 cm^{-1} (stretching modes). An integrated absorbance in those regions may be used but can be susceptible to edge distortions due to molecularly non-specific scattering.[24] While multiplicative scatter correction[25] and rigorous optical theory[26,27] approaches are emerging, an approximation to removing scattering distortion is the use of second derivatives of spectra.[28] The sum of the absolute values of the second derivative data is then indicative of the overall chemical composition of the tissue. The Savitzky-Golay filter used for computing derivatives also reduces noise while preserving peak heights and widths, providing a high SNR **I** (Fig. 2) that captures features from important spectral bands. Yet another alternative is to calculate the Gram-Schmidt intensity of the interferogram of the sample,[29] which could be a faster route by precluding the FT-process. The image, however, would retain both structural and biochemical contributions from all functional groups and scattering interfaces. Finally, another approach could be to use the bright field optical microscopy image. The optical image, however, may not contain sufficient contrast, have differences observed in the IR image or may experience a mismatch in resolution. The IR "bright field" equivalent, which is simply the height of the centerburst may be used. Since a background is collected for absorbance data, the sample data set can be easily corrected for illumination differences. This approach can be considered a combination of both IR spectral (absorbance) and visible optical (scattering) imaging.

Having chosen an 'ideal' image **I**, its edge map $\mathbf{E}_\mathrm{I}$ is computed. Next, each eigenimage is filtered using the same kernel as that used for the ideal image edge map and edge maps $\mathbf{E}_\mathrm{j}, j = 1, \ldots, M$ can be found. In practice, the number of significant eigenimages are much smaller than the number of spectral data points. Hence, it is prudent to consider a smaller subset of eigenimages to save computation time. In carefully examining resulting images from the MNF transform, we noticed that most information content was in the first 30 eigenimages. Hence, we chose to examine a smaller subset of MNF transformed eigenimages ($P = 64$) from the available spectral data points (1640). This represents a substantial reduction in the time for comparison and data storage needed. Next, the root mean square error (RMSE) between $\mathbf{E}_\mathrm{I}$ and $\mathbf{E}_\mathrm{j}, j = 1, 2, 3 \ldots, P$ is computed as a measure of the spectral similarity of the images using

$$RMSE_j = \sqrt{\sum_{p=1}^{b_x} \sum_{q=1}^{b_y} \left(E_I(p,q) - E_j(p,q)\right)^2} \; j = 1, 2, 3, \ldots P \quad (6)$$

Where, bx and by are the pixels along the row and columns of the image array. A typical plot of RMSE as a function of eigenimage number is shown in Fig. 3. The RMSE prior to sorting is the decreasing order of importance from the MNF transform and shows that factors corresponding to higher eigenvalues (lower eigneimage number) may not necessarily have significant features. While the eigenvalue curve is obviously monotonically decreasing, the RMSE curve of the MNF eigenimages displays significant fluctuation. Re-ordering by RMSE values not only re-orders the important eigenimages by assigning them a lower number but makes the curve smooth and amenable to accurately determining saturation or calculating derivatives. It is notable, though, that the actual RMSE error is not affected by re-ordering and the information content of all the transformed data is only re-prioritized and not altered in any manner. It is instructive to compare the RMSE ordered and MNF ordered eigenimages (Fig. 3). While it appears that ~60 eigenimages would be important from the MNF-ordered plot, the RMSE-ordered curve indicates that~30 images may be useful in the inverse transform. This discordance is due to the sensitivity to any structure in the image in MNF-ordered data, while the RMSE-ordered data are only sensitive to the structure in the reference image. Hence, prioritizing eigenimages by RMSE is likely beneficial. It should be noted that the eigenimage number for the unsorted RMSE and MNF-order is the same. While there is a generally increasing trend (decreasing importance) for both values, the RMSE plot appears to be noisy. Since the RMSE is directly a measure of concordance, hence, we sorted eigenimage numbers based on increasing RMSE and assigned them new eigenimage numbers based on RMSE values.

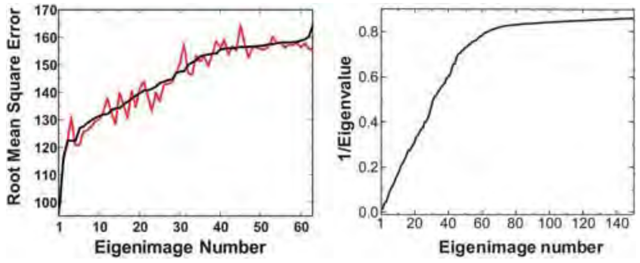


Fig. 3 (Left) Typical error plot before sorting (red) and after sorting (black) for RMSE, where the increasing eigneimage number indicates a decreasing order of importance. (Right) Eigenimage order (number) as ranked by the MNF transform.

The importance of the alteration by sorting can be understood by examining the RMSE plot (Fig. 3) in conjunction with the spatial features in eigenimages as seen in Fig. 4. Eigenimages and their corresponding edge maps demonstrate first that images with significant features (*e.g.* numbers 1, 3, 10 and 18) have well defined edge maps while those without significant features (*e.g.* number 46) have nondescript edge maps. Second, the spatial similarity of early factors with the reference edge map results in lower RMSE values that increase with increasing noise. When information content of the image is dominated by noise, the RMSE between any edge map and that of the ideal image is nearly independent of the actual edges, resulting in the plateau region of the RMSE curve. Eigenimages close to the chosen cut off have edge maps with a semblance of features buried in noise. By choosing all factors corresponding to RMSE values less than that at the cut-off point, we select only those factors with significant features. The derivative of the curve in the plateau region is negligible and could also be utilized in finding the cut-off point. We choose the cutoff to be the point after which the derivative does not rise more than μ + 3σ, where μ and σ correspond to the mean and standard deviation of the derivative of flat region of the curve. This is a very strict condition which maintains a high degree of spectral detail. Other cutoff values may be chosen, for example, μ + σ or simply the first image whose RMSE exceeds μ. Our interest was in preserving as much spectral detail as possible; hence, we adopt a criterion that may be more stringent than most and likely represents a lower level of improvement in SNR than other cutoffs. In summary, computing the MNF transform, selecting eigenimages based on sorted RMSE from edge maps and computing the inverse MNF using the reduced set of eigenimages prior to the cutoff is a completely automated noise reduction algorithm that does not require human input. There are choices that can be made while setting up the protocol, for example, in choice of the reference image, that are under operator control. Once the protocol is finalized, however, the process is entirely automated and can be high throughput. Thus, the criteria of both objectivity and automation for noise reduction are addressed.

3. Experimental

Tissue used for this study (Biomax Inc.) was processed as per procedures reported earlier.[30] Spectroscopic imaging data are acquired using the Perkin-Elmer Spotlight 400 imaging spectrometer that is equipped with a linear array detector and samples a 6.25μm × 6.25μm area per pixel. An undersampling ratio of two with reference to the He-Ne laser and mirror scanning speed of 1 cm/s is used to sample the interferogram to provide a spectral resolution of 4 cm^{-1}. The interferogram at every pixel is then Fourier transformed using a zero-filling factor of two and N-B medium apodization and truncated to

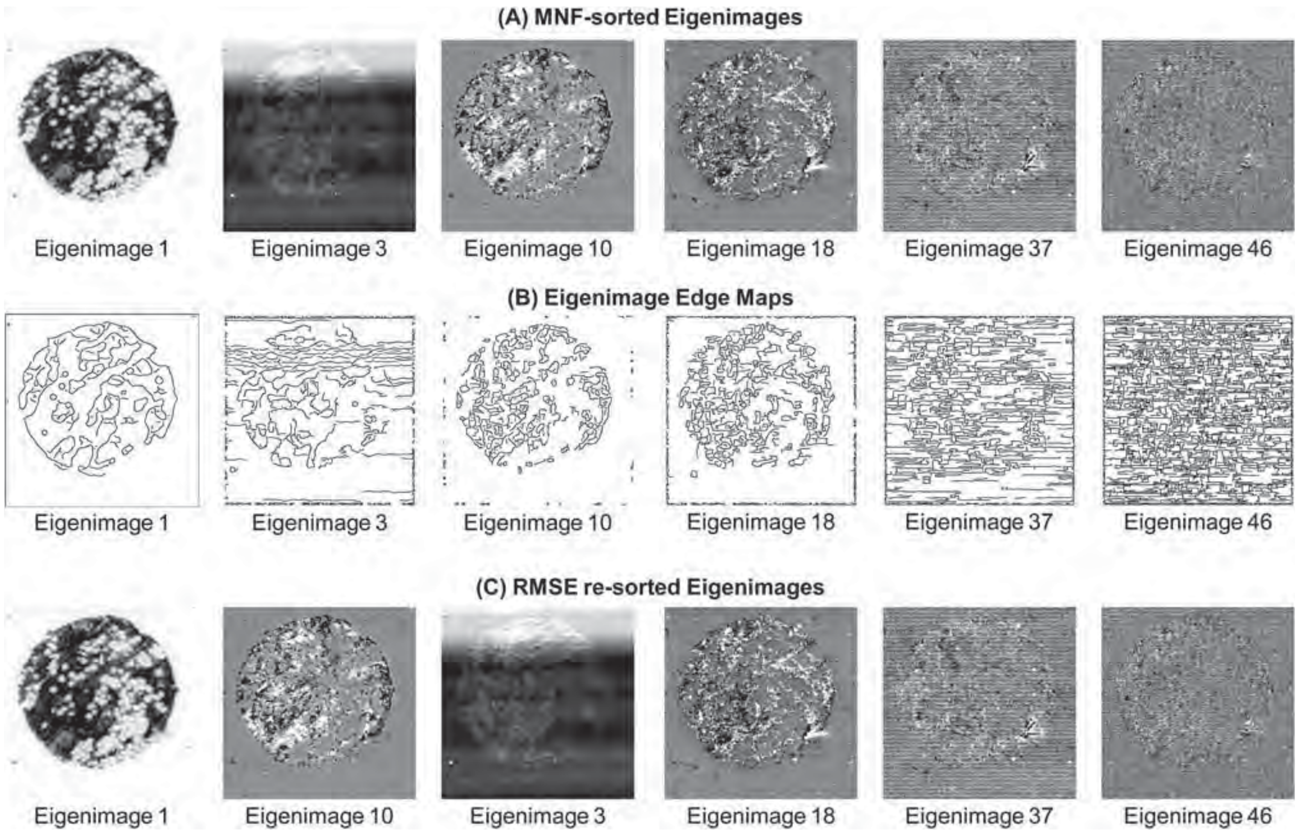


Fig. 4 Typical eigenimages (A) ordered by the MNF transform. (B) corresponding edge maps for MNF-ordered images. (C) RMSE re-sorted eigenimages within this subset.

4000−720 cm^{-1} for efficient storage. A background single beam reference is collected by averaging 120 scans and sample spectra are acquired by averaging two interferometer scans. To validate the method for different instruments, we implemented the same algorithm on data acquired from a system equipped with a large two-dimensional array detector (Varian Stingray). The system consists of a Varian 7000 Spectrometer coupled to a microscope accessory, UMA-400. The imaging detector is a liquid nitrogen-cooled mercury cadmium telluride (MCT) focal plane array that is windowed to 32 × 32 elements (Santa Barbara Focalplane). The detector samples an area of 175μm × 175μm at the sample. Interferograms were acquired in rapid scan mode with an undersampling ratio of 2 at a spectral resolution of 4 cm^{-1} and Fourier transformed using a factor of two zero-filling and Norton-Beer(NB) medium apodization. The data were truncated to 4000 − 950 cm^{-1} for storage. For these data, the number of co-additions were varied (1, 2, 4, 8, 16, 32 and 64 scans) to obtain a range of poor to good SNR data. The background reference was collected at 120 co-additions.

All software used was written in-house or utilized programs in ENVI/IDL. Computing MNF transforms involves estimating noise statistics. ENVI can use a shift difference method to compute noise statistics, which assumes that every pixel contains both signal and noise, and that adjacent pixels contain the same signal but different noise. A shift difference is performed on the data by differencing adjacent pixel above and to the right of each pixel and averaging the results to obtain the 'noise' value to assign to the pixel being processed. To the extent that this assumption is not true, the noise statistics estimate is in error. Rigorously, the noise should be estimated using repeat measurements, as that is easily possible in FT-IR imaging. With the commercial raster scanning system, however, we were unable to obtain successive measurements without a new scan. The positioning error on the stage was such that slight pixel shifting was observed, precluding true averaging at every pixel. Hence, we employed the shift difference method in this study. The pixel size being set smaller than the lowest resolution achievable, and the general nature of large phases in the data here likely result in the estimate being close.

4. Results and discussion

In order to quantify the SNR gain from noise reduction, we first acquired high SNR data using the linear array system as a base for simulations and as a comparator. Poor SNR data is simulated from this data by adding noise from a normal distribution with different standard deviations (σ = 0.001, 0.01, 0.1 and 0.4 a.u.) as shown for a single pixel in Fig. 5(A). Resulting spectra after noise reduction are shown in Fig. 5(B). An improvement is apparent, even in cases where noise appears to overwhelm spectral features. We then acquired data on the large array system in which single scan acquisition was compared to 64 scan acquisition (~70 fold slower). As expected, noise-reduced data were found to be comparable to the high scan numbers. To

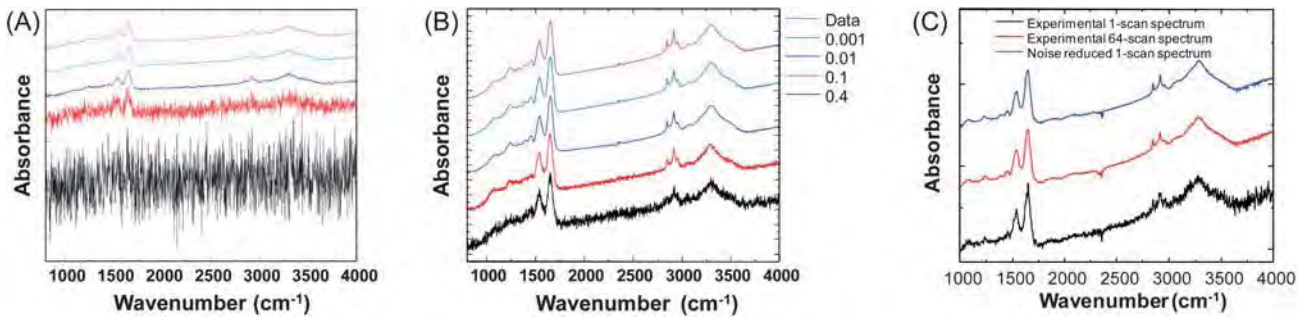


Fig. 5 (A) Acquired high SNR data and simulated noisy spectra obtained by adding noise ($\sigma = 0.001, 0.01, 0.1$ and 0.4 a.u.), showing the degradation in data quality. Spectra are offset for clarity. (B) Corresponding spectra after noise reduction. (C) Absorption spectrum (1-scan in black) compared to the resulting spectrum from the same pixel after noise reduction (blue) and to that acquired by averaging 64 scans (red).

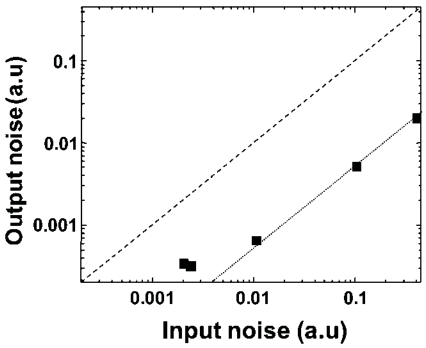


Fig. 6 Noise before (input noise) and after application of the algorithm (output noise). An order of magnitude improvement can be observed.

quantify the benefits of post-acquisition processing, the reduction in noise achieved is quantified in Fig. 6. Noise values were calculated using the non-absorbing 1950 cm^{-1} –2000 cm^{-1} region with 41 spectral points around 1975 cm^{-1} and are averages of 1024 spectra.

The dashed diagonal is the unity gain line that separates decrease or increase in noise upon application of the algorithm. The plot indicates success in applicability over three orders of magnitude of input noise where an order of magnitude noise reduction is observed. The actual noise reduction depends on the number of factors chosen for the inverse transform, the number of pixels in the original data set and the degree of correlation in the noise. If the noise is high enough, the benefit is observed to be proportional to the input noise. For very low noise cases, the plot indicates that it becomes difficult to improve the data further. This behavior likely arises from the distribution of noise and information in eigenimages. It must be noted that many of the eigenimages rejected in the inverse transform do contain information and all selected do contain noise that is both correlated and uncorrelated. Hence, the limitation of the process arises from both correlated noise and the need to balance information content of eigenimages with the opportunity to reduce noise, We have used a fairly conservative approach to noise reduction in that fewer eigenimages could have been selected in the inversion, which may also explain the lack of significant improvements when the input noise is low. It is interesting to note that a previous application of the MNF transform[10] also provided a limit to the improvement possible with this approach, but in the high noise limit. There, the high input noise data were found to contain a low frequency response in the spectra of inverse transformed data that limited the noise reduction achieved. In summary, the forward-reverse transform approach appears to be bounded in its ability to improve data quality in both the high noise (as previously shown) and low noise cases (as observed here). These limits must be considered when designing data acquisition protocols that take advantage of this post-processing approach.

From the trading rules of FT-IR spectroscopy,[3,31] a factor of *n* improvement in SNR requires an increase of n^2 in data acquisition time. Hence, a method to increase data acquisition rate without loss in its quality could involve rapid data collection at a low SNR followed by application of numerical techniques for noise reduction. The order of magnitude improvement, as we show above, allows for close to two orders of magnitude reduction in scanning time. To test this hypothesis, we compared noise reduced data from a single interferometer scan with data obtained by averaging 64 scans (Fig. 5(D)). Spectra with only one scan, after noise reduction, closely resemble spectra obtained from 64 scans experimentally. Caution must be exercised, however, in claiming that mathematical techniques provide precisely equivalent data. As can be seen from the spectra, there are some low frequency noise components in the noise–reduced spectrum that were not eliminated.[32] Noise reduction has important implications in areas where data quality cannot be improved by averaging (*e.g.* kinetics measurements),[33] for low-throughput configurations such as total internal reflection sampling,[34,35,36] where large quantities of data are acquired or where the analyte signal is low. An interesting test case in to perform histopathology without human intervention[37] faster than with current data acquisition protocols. Briefly, FT-IR microspectroscopy combined with pattern recognition tools[38] is rapidly developing as a potential tool for automated structure[39] and disease recognition[40,41,42] within complex tissue by a number of groups.[43,44] Unfortunately, the time to acquire data from large numbers of samples is prohibitive. For example, a recent study[30] reported the quantitative evaluation of classification using large sample and data sets that required many months to acquire. Reducing data acquisition time through automated noise reduction will help reduce time in laboratory studies. When the approach is translated to clinical venues, it will serve to enhance the speeds and throughput of samples. As an example, Fig. 7 illustrates the benefits of using automated noise reduction.

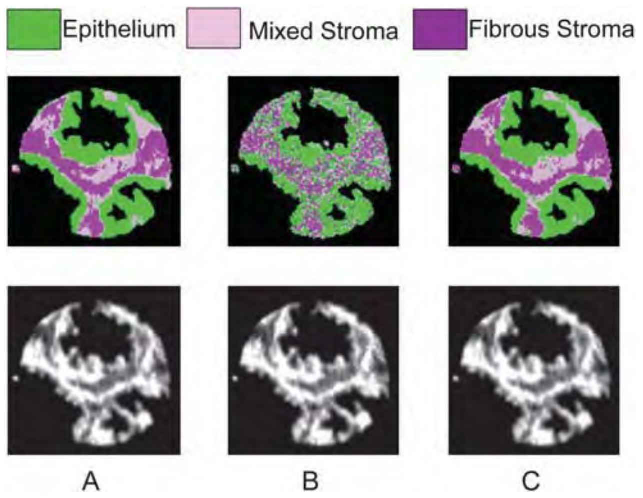


Fig. 7 Effect of automated noise reduction on prostate tissue classification. Top row: classification results, Bottom row: absorbance in a tissue sample at 1080 cm^{-1} (A) high SNR data in which measured baseline noise is ~0.001 a.u. with the corresponding classified images showing three types of cells. (B) Lower SNR data in which measured baseline noise is ~0.005 a.u., demonstrating that the classification becomes noisy and (C) noise reduced data set obtained from the data in (B), demonstrating that the classification errors are reduced.

Prostate tissue is classified into its constituent cell types. Classification is inaccurate for the higher noise case but is recovered when the noise is reduced. The time for data acquisition for this 500 μm × 500 μm image set was reduced from ~45 min to less than 2 min. While the result demonstrates qualitative agreement between the classified images, we examine next a detailed quantitative assessment of the fidelity of inverse transformed data and the benefits of noise rejection for tissue classification.

The accuracy of tissue classification is related to SNR of the data, as has been demonstrated previously for prostate tissue.[45] Here, we sought to apply the same exercise to breast tissue. We acquired data from 10 tissue samples consisting of almost 8000 spectra per sample. The samples contain a variety of cell types and disease states. As a first step towards classification for disease diagnoses, breast tissue is divided into two cell types (epithelial cells, which are indicated in green, and stromal cells, which are indicated in magenta). The effect of decreasing data quality can be seen in classified images shown in Fig. 8A–D (top row). Noise in the underlying absorbance data increases from A to D, thereby noise in the classified images increases progressively until all ability to segment tissue is lost for noise levels ~0.1 a.u (Fig. 8D). We quantified classification accuracy, further as measured by calculating the area under the curve (AUC) of the receiver operating curve(ROC)[46] for pixels that meet the threshold for classification, in Fig. 8 (E). As a function of average noise in the absorbance data, AUC values finally fall to about 0.5, which is equivalent to random guessing and does not provide any useful classification information. At the higher noise levels, some tissue pixels are not even recognized as meeting the threshold for inclusion. For intermediate noise levels, classification accuracy decreases.

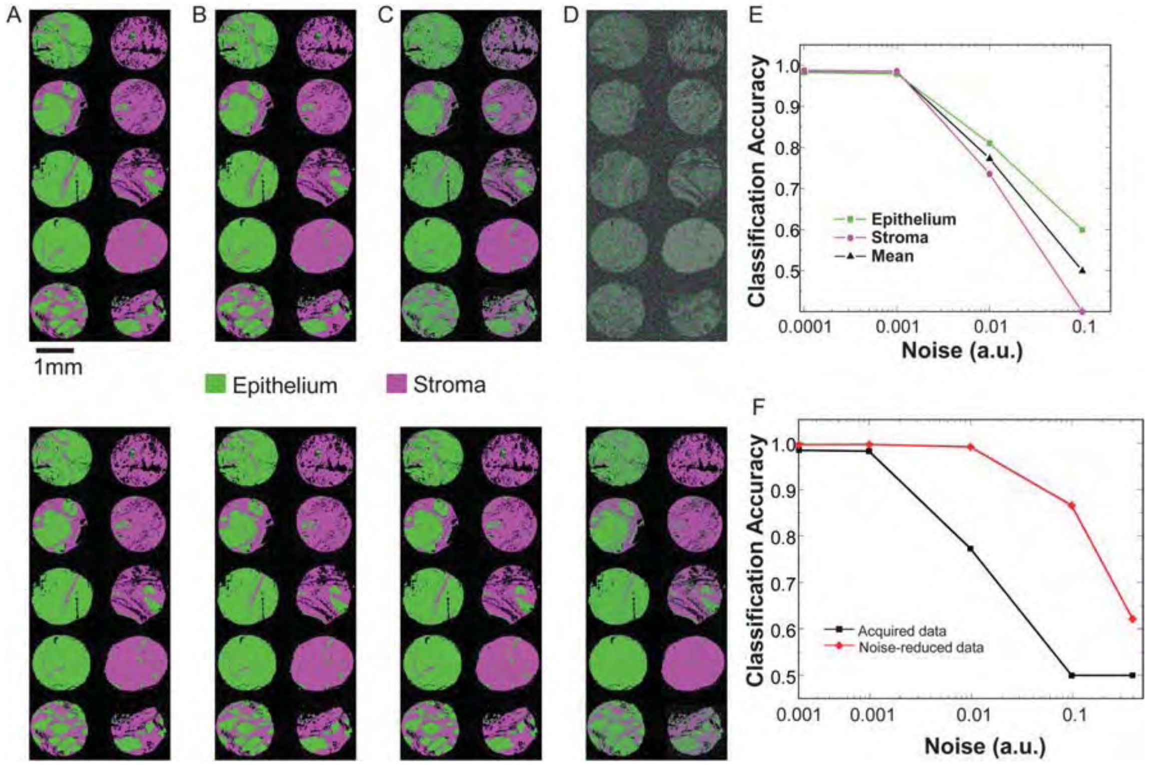


Fig. 8 Effect of noise in the absorbance data on image classification is illustrated for breast tissue in A–D (top panel), where the noise in the data is calculated to be 0.0001, 0.001, 0.001 and 0.1 a.u., respectively. (E) Classification accuracy, as measured by the area under the receiver operating characteristic curve, decreases with increasing noise for both cell types. Image classification is shown upon using the noise reduction algorithm (A–D, bottom panel). (F) Classification accuracy before and after noise reduction.

The impact of noise reduction on classification is demonstrated in the bottom panel of Fig. 8. Classified images for each noise-reduced case (A–D) demonstrate that classification accuracy for all cases appear to be comparable to classification accuracy for low noise cases. Examination of classified images and classification accuracy values indicates that noise reduction improves classifier performance in each case. For low noise data, noise reduction does not appear to significantly impact classification since the classification accuracy is almost 100%. On the other hand, noise reduction significantly improves classification from FT-IR spectroscopic imaging data with higher noise levels. Hence, a potential route to faster data acquisition for histopathology, without the need to modify hardware or change any experimental configuration, can be proposed based on post-processing noise reduction. The ten-fold increase in noise of the data to provide the same classification accuracy implies that $\sim$100-fold decrease in data acquisition time may be obtained. Instead of requiring $\sim$300 h (12 days) to scan a 1 cm × 1 cm area with a large focal plane detector, the proposed approach will allow the same in $\sim$3 h. This conclusion is one of the more important aspects of this study, implying that a careful noise rejection protocol can speed up data acquisition to make present FT-IR imaging instrumentation perform analyses within clinically acceptable time periods.

5. Conclusions

An objective eigenimage selection scheme based on structural features has been proposed here for automated noise reduction after data acquisition. An order of magnitude reduction in noise could be achieved using this algorithm when the noise was not very low. Applied to obtaining results from samples, for example for tissue classification, there is an equivalent recovery of correct results at higher noise levels. The improvement translates directly into a reduction in time required for data collection. It must be noted that the gain here is through post-acquisition computational techniques and does not involve changes in instrumentation hardware or data acquisition schemes. Hence, it is easy to implement and inexpensive to deploy. It is anticipated that the automated nature of the proposed approach will allow it to become routinely applied to enhance data quality and the recover scientific results with lower experimental efforts (time, expense and hardware) in data acquisition.

6. Acknowledgements

The authors thank Frances Nell Pounder for acquiring breast tissue data. Funding for this work was provided in part by the Department of Defense Prostate Cancer Research Program, by the National Center for Supercomputing Applications and the University of Illinois, under the auspices of the NCSA/UIUC faculty fellows program and Susan G. Komen for the Cure.

References

1 E. N. Lewis, P. J. Treado, R. C. Reeder, G. M. Story, A. E. Dowrey, C. Marcott and I. Levin, *Anal. Chem.*, 1995, **67**, 3377–3381.
2 P. R. Griffiths, *Anal. Chem.*, 1972, **44**, 1909–1913.
3 C. M. Snively and J. L. Koenig, *Appl. Spectrosc.*, 1999, **53**, 170–177.
4 R. Bhargava and I. W. Levin, *Anal. Chem.*, 2001, **73**, 5157–5167.
5 C. M. Snively, S. Katzenberger, G. Oskarsdottir and J. Lauterbach, *Opt. Lett.*, 1999, **24**, 1841–1843.
6 S. W. Huffman, R. Bhargava and I. W. Levin, *Appl. Spectrosc.*, 2002, **56**, 965–969.
7 R. Bhargava, T. Ribar and J. L. Koenig, *Appl. Spectrosc.*, 1999, **53**, 1313–1322.
8 A. Green, M. Berman, P. Switzer and M. Craig, *IEEE Trans. Geosci. Remote Sens.*, 1988, **26**, 65–74.
9 J. B. Lee, A. S. Woodyatt and M. Berman, *IEEE Trans. Geosci. Remote Sens.*, 1990, **28**, 295–304.
10 R. Bhargava, S. Wang and J. L. Koenig, *Appl. Spectrosc.*, 2000, **54**, 486–495.
11 M. J. Wabomba, Y. Sulub and G. W. Small, *Appl. Spectrosc.*, 2007, **61**, 349–358.
12 F. Vogt, *Curr. Anal. Chem.*, 2006, **2**, 107–127.
13 F. Vogt, J. Cramer and K. Booksh, *J. Chemom.*, 2005, **19**, 510–520.
14 R. Bhargava, S. Wang and J. L. Koenig, *Appl. Spectrosc.*, 2000, **54**, 1690–1706.
15 J. Boardman and F. Kruse, *Proc. ERIM Tenth Thematic Conf. Geo. Remote Sens.*, 1994, 407–418.
16 P. Wentzell, D. Andrews, D. Hamilton, K. Faber and B. Kowalski, *J. Chemom.*, 1997, **11**, 339–366.
17 S. Qin and R. Dunia, *J. Process Control*, 2000, **10**, 245–250.
18 R. B. Cattell, *Multivar. Behav. Res.*, 1966, **1**, 245–276.
19 S. Wold, *Technometrics*, 1978, **20**, 397–405.
20 S. Valle, W. Li and S. Qin, *Ind. Eng. Chem. Res.*, 1999, **38**, 4389–4401.
21 F. N. Pounder, R. K. Reddy, R. Bhargava, Submitted (2010).
22 R. Gonzalez, R. Woods, S. Eddins, *Digital image processing using MATLAB*, Prentice Hall, USA, 2003, pp. 378–425.
23 J. Canny, *IEEE T. Pattern Anal.*, 1986, **8**, 679–698.
24 R. Bhargava, S. Wang and J. L. Koenig, *Appl. Spectrosc.*, 1998, **52**, 323–328.
25 P. Bassan, A. Kohler, H. Martens, J. Lee, E. Jackson, N. Lockyer, P. Dumas, M. Brown, N. Clarke and P. Gardner, *J Biophotonics*, Apr 22, 2041, **2010**.
26 B. J. Davis, P. S. Carney and R. Bhargava, *Anal. Chem.*, 2010, **82**, 3487–3499.
27 B. J. Davis, P. S. Carney and R. Bhargava, *Anal. Chem.*, 2010, **82**, 3474–3486.
28 A. Savitzky and M. Golay, *Anal. Chem.*, 1964, **36**, 1627–1639.
29 R. Bhargava and I. W. Levin, *Appl. Spectrosc.*, 2004, **58**, 995–1000.
30 D. C. Fernandez, R. Bhargava, S. M. Hewitt and I. W. Levin, *Nat. Biotechnol.*, 2005, **23**, 469–474.
31 P. R. Griffiths, J. A. De Haseth, *Fourier Transform Infrared Spectrometry*, John Wiley & Sons, New York, USA, 1986, 2nd edn, pp. 254–260.
32 B. U. Dongsheng, S. W. Huffman, J. A. Seelenbinder and C. W. Brown, *Appl. Spectrosc.*, 2005, **59**, 575–583.
33 R. J. Hendershot, P. T. Fanson, C. M. Snively and J. A. Lauterbach, *Angew. Chem., Int. Ed.*, 2003, **42**, 1152–1155.
34 A. J. Sommer, L. G. Tisinger, C. Marcott and G. M. Story, *Appl. Spectrosc.*, 2001, **55**, 252–256.
35 K. L. A. Chan and S. G. Kazarian, *Appl. Spectrosc.*, 2003, **57**, 381–389.
36 B. M. Patterson and G. J. Havrilla, *Appl. Spectrosc.*, 2006, **60**, 1256–1266.
37 G. Srinivasan and R. Bhargava, *Spectroscopy*, 2007, **22**, 30–43.
38 P. Lasch and D. Naumann, *Cell. Mol. Biol.*, 1998, **44**, 189–202.
39 R. Mendelsohn, E. P. Paschalis and A. L. Boskey, *J. Biomed. Opt.*, 1999, **4**, 14–21.
40 C. Petibois and G. Déléris, *Trends Biotechnol.*, 2006, **24**, 455–462.
41 R. K. Sahu, S. Argov, A. Salman, U. Zelig, M. Huleihel, N. Grossman, J. Gopas, J. Kapelushnik and S. Mordechai, *J. Biomed. Opt.*, 2005, **10**, 054017–10.
42 C. Krafft, L. Shapoval, S. B. Sobottka, K. D. Geiger, G. Schackert and R. Salzer, *Biochim. Biophys. Acta, Biomembr.*, 2006, **1758**, 883–891.
43 M. Diem, M. Romeo, S. Boydston-White, M. Miljković and C. Matthaus, *Analyst*, 2004, **129**, 880–885.
44 *Vibrational Spectroscopy for Medical Diagnosis*, ed. M. Diem, P. R. Griffiths, J. M. Chalmers, John Wiley and Sons, 2008.
45 R. Bhargava, *Anal. Bioanal. Chem.*, 2007, **389**(4), 1155–69, Epub 2007 Sep 5.
46 J. A. Hanley and B. J. McNeil, *Radiology*, 1982, **143**, 29–36.

ARTICLE IN PRESS

Information Sciences xxx (2010) xxx–xxx

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Repairing fractures between data using genetic programming-based feature extraction: A case study in cancer diagnosis

Jose G. Moreno-Torres [a,*], Xavier Llorà [b], David E. Goldberg [c], Rohit Bhargava [d]

[a] Department of Computer Science and Artificial Intelligence, Universidad de Granada, 18071 Granada, Spain
[b] National Center for Supercomputing Applications (NCSA), University of Illinois at Urbana-Champaign 1205 W. Clark Street, Urbana, Illinois, USA
[c] Illinois Genetic Algorithms Laboratory (IlliGAL) University of Illinois at Urbana-Champaign 104 S. Mathews Ave, Urbana, Illinois, USA
[d] Department of Bioengineering, University of Illinois at Urbana-Champaign 405 N. Mathews Ave, Urbana, Illinois, USA

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Genetic programming
Feature extraction
Fractures between data
Biological data
Cancer diagnosis
Different laboratories

ABSTRACT

There is an underlying assumption on most model building processes: given a learned classifier, it should be usable to explain unseen data from the same given problem. Despite this seemingly reasonable assumption, when dealing with biological data it tends to fail; where classifiers built out of data generated using the same protocols in two different laboratories can lead to two different, non-interchangeable, classifiers. There are usually too many uncontrollable variables in the process of generating data in the lab and biological variations, and small differences can lead to very different data distributions, with a fracture between data.

This paper presents a genetics-based machine learning approach that performs feature extraction on data from a lab to help increase the classification performance of an existing classifier that was built using the data from a different laboratory which uses the same protocols, while learning about the shape of the fractures between data that motivated the bad behavior.

The experimental analysis over benchmark problems together with a real-world problem on prostate cancer diagnosis show the good behavior of the proposed algorithm.

1. Introduction

The assumption that a properly trained classifier will be able to predict the behavior of unseen data from the same problem is at the core of any automatic classification process. However, this hypothesis tends to prove unreliable when dealing with biological (or other experimental sciences) data, especially when such data is provided by more than one laboratory, even if they are following the same protocols to obtain it.

The specific problem this paper attempts to solve is the following: we have data from one laboratory (dataset A), and derive a classifier from it that can predict its category accurately. We are then presented with data from a second laboratory (dataset B). This second dataset is not accurately predicted by the classifier we had previously built due to a fracture between the data of both laboratories. We intend to find a transformation of dataset B (dataset S) where the classifier works.

Evolutionary computing, as introduced by Holland [27]; is based on the idea of the survival of the fittest, evoked by the natural evolutionary process. In genetic algorithms (GAs) [21], solutions (genes) are more likely to reproduce the fitter

* Corresponding author. Tel.: +34588998183.
E-mail addresses: jose.garcia.mt@decsai.ugr.es (J.G. Moreno-Torres), xllora@illinois.edu (X. Llorà), deg@illinois.edu (D.E. Goldberg), rbx@uiuc.edu (R. Bhargava).

0020-0255/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2010.09.018

they are, and random sporadic mutations help maintain population diversity. Genetic Programming (GP) [33] is a development of those techniques, and follows a similar pattern to evolve tree-shaped solutions using variable-length chromosomes.

Feature extraction, as defined by Wyse et al. [56], 'consists of the extraction a set of new features from the original features through some functional mapping'. Our approach to the problem can be seen as feature extraction, since we build a new set of features which are functions of the old ones. However, we have a different goal than that of classical feature extraction, since our intention is to fit a dataset to an already existing classifier, not to improve the performance of a future one.

In this work, we intend to demonstrate the use of GP-based feature extraction to unveil transformations in order to improve the accuracy of a previously built classifier, by performing feature extraction on a dataset where said classifier should, in principle, work; but where it does not perform accurately enough. We test our algorithm first on artificially-built problems (where we apply ad hoc transformations to datasets from which a classifier has been built, and use the dataset resulting from those transformations as our problem dataset); and then on a real-world application where biological data from two different medical laboratories regarding prostate cancer diagnosis are used as datasets A and B.

Even though the method proposed in this paper does not attempt to reduce the number of features or instances in the dataset, it can still be regarded as a form of data reduction because it unifies the data distributions of two datasets; which results in the capability of applying the same classifier to both of them, instead of needing two different classifiers, one for each dataset.

The remainder of this paper is organized as follows: in Section 2, some preliminaries about the techniques used and some approaches to similar problems in the literature are presented. Section 3 details the real-world biological problem that motivates this paper. Section 4 has a description of the proposed algorithm GP-RFD; and Section 5 includes the experimental setup, along with the results obtained, and an analysis. Finally, in Section 6 some concluding remarks are made.

2. Preliminaries

This section is divided in the following way: in Subsection 2.1 we introduce the notation that has been used in this paper. Then we include an introduction to GP in Subsection 2.2, a brief summary of what has been done in feature extraction in Subsection 2.3, and a short review of the different approaches we found in the specialized literature on the use of GP for feature extraction in Subsection 2.4. We conclude mentioning some works related to the finding and repair of fractures between data in Subsection 2.5.

2.1. Notation

A classification problem is considered with:

- A set of input variables $X = \{x_i / i = 1,\ldots,n_v\}$, where n_v is the number of features (attributes) of the problem.
- A set of values for the target variable (class) $C = \{C^j / j = \{1,\ldots,n_c\}\}$, where n_c is the number of different values for the class variable.
- A set of examples $E = \{e^h = (e^h_1,\ldots,e^h_{n_v},C^h)/h = 1,\ldots,n_e\}$, where C^h is the class label for the sample e^h, and n_e is the number of examples.

When describing the problem, we mention datasets A, B and S. They correspond to:

- A: the original dataset that was used to build the classifier.
- B: the problem dataset. The classifier is not accurate on this dataset, and that is what the proposed algorithm attempts to solve.
- S: the solution dataset, result of applying the evolved transformation to the samples in dataset B. The goal is to have the classifier performance be as high as possible on this dataset.

When performing experiments and obtaining the evolved expressions, we use the following notation: when artificially creating a dataset B by means of a fabricated transformation over dataset A, we have $B = \{b_i / i = 1,\ldots,n_v\}$ be the attributes in dataset B and $A = \{a_i / i = 1,\ldots,n_v\}$ be the ones from dataset A. In appendix A, we show the learned transformations for the prostate cancer problem. The attributes shown are those corresponding to dataset S, and are represented as $S = \{s_i / i = 1,\ldots,n_v\}$.

2.2. Genetic programming

A GA [21] is a stochastic optimization technique inspired by nature's development of useful characters. It is based on the idea of survival of the fittest [11] in the following way: given a population of possible solutions to a problem (represented by chromosomes), there is some selection procedure that favors the fitter ones (i.e., the ones that provide a higher-quality solution); and the selected chromosomes get an opportunity to pass down their genetic material to the next generation via some crossover operator; which usually builds new individuals from the combination of old ones. In some variations of the algorithm, random mutations are sporadically introduced to help maintain biological diversity in the population.

GP, as proposed by John Koza in 1992 [33], uses a selectorecombinative schema where the solutions are represented by trees; which are encoded as variable-length chromosomes. It was originally designed to automatically develop programs, but it has been used for multiple purposes due to its high expressive power and flexibility. In the words of Poli and Langdon [46], 'GP is a systematic, domain-independent method for getting computers to solve problems automatically starting from a high-level statement of what needs to be done. Using ideas from natural evolution, GP starts from an ooze of random computer programs, and progressively refines them through processes of mutation and sexual recombination, until solutions emerge. This is all done without the user having to know or specify the form or structure of solutions in advance. GP has generated a plethora of human-competitive results and applications, including novel scientific discoveries and patentable inventions'.

There are a few details about GP that make it different from standard GAs:

- Crossover: the most commonly used operator is one-point crossover, which is analogous to the GA classical one, but where subtrees instead of a specific gene signal where the cut is made.
- Even though mutation was used in the early literature regarding the evolution of programs (see [7,10,16]) Koza chose not to use it [33,34], as he wished to demonstrate that mutation was not necessary. This has significantly influenced the field, and mutation was often omitted from GP runs. However, mutation has proved useful since then (see [5,42], for example); and its use is widely spread nowadays. Multiple different mutation operators have been proposed in the literature [44].
- Treatment of constants: the discovery of constants is one of the hardest issues in GP. Koza proposed a solution called Ephemeral Random Constant (ERC), which uses a fixed terminal (e) to represent a constant. The first time one of such constants is evaluated, it gets assigned a random value. From there on, it retains that value throughout the whole run. A number of alternatives have been proposed in the literature [14,49], but ERC remains the most used one.
- Automatically defined functions: ADFs were also first proposed by Koza [34]. The idea is to permit each individual to evolve more than one tree simultaneously; having the extra trees work as primitives that can be called from the main one.

GP has been applied often to classification [13]. Among the latest advances in the field, we would like to mention those dedicated to high dimensional problems [35,6], variations in population size [31,32], and applications to other related fields [58,3].

2.3. Feature extraction

Feature extraction creates new features as functional mappings of the old ones. It has been used both as a form of preprocessing, which is the approach we use in this paper, and also embedded with a learning process in wrapper techniques. An early proposer of such a term was probably Wyse in 1980, in a paper about intrinsic dimensionality estimation [56]. There are multiple techniques that have been applied to feature extraction throughout the years, ranging from principal component analysis to support vector machines to GAs (see [28,45,43], respectively, for some examples).

Among the foundations papers in the literature, Liu's book in 1998 [38] is one of the earlier compilations of the field. As a result of a workshop held in 2003 [24], Guyon and Elisseeff published a book with an important treatment of the foundations [25].

2.4. Genetic programming-based feature extraction

GP has been used extensively to optimize feature extraction and selection tasks. One of the first contributions in this line was the one published by Tackett in 1993 [53], who applied GP to feature discovery and image discrimination tasks.

We can consider two main branches in the philosophy of GP-based feature extraction:

On one hand, we have the proposals that focus only on the feature extraction procedure, of which there are multiple examples: Sherrah et al. [50] presented in 1997 the evolutionary pre-processor (EPrep), which searches for an optimal feature extractor by minimizing the misclassification error over three randomly selected classifiers. Kotani et al.'s work from 1999 [30] determined the optimal polynomial combinations of raw features to pass to a k-nearest neighbor classifier. In 2001, Bot [8] evolved transformed features, one-at-a-time, again for a k-NN classifier, utilizing each new feature only if it improved the overall classification performance. Zhang and Rockett, in 2006, [61] used multiobjective GP to learn optimal feature extraction in order to fold the high-dimensional pattern vector to a one-dimensional decision space where the classification would be trivial. Lastly, also in 2006, Guo and Nandi [23] optimized a modified Fisher discriminant using GP, and then Zhang et al. extended their work by using a multiobjective approach to prevent tree bloat [62], and applied a similar method to spam filtering [60].

On the other hand, some authors have chosen to evolve a full classifier with an embedded feature extraction step. As an example, Harris [26] proposed in 1997 a co-evolutionary strategy involving the simultaneous evolution of the feature extraction procedure along with a classifier. More recently, Smith and Bull [52] developed a hybrid feature construction and selection method using GP together with a GA. FLGP, by Yin et al. [37] is yet another example, where 'new features extracted by certain layer are used to be the training set of next layer's populations'.

2.5. Finding and repairing fractures between data

Throughout the literature there have been a number proposals to quantify the amount of dataset shift (in other words, the size of the fracture in the data). This shift is usually due to time passing (the data comes from the same source at a latter time), but it can also be due to the data being originated by different sources, as is the case in this paper. Some of the most relevant works in the field are: Wang et al. [54], where the authors present the idea of correspondence tracing. They propose an algorithm for the discovering of changes of classification characteristics, which is based on the comparison between two rule-based classifiers, one built from each dataset. Yang et al. [57] presented in 2008 the idea of conceptual equivalence as a method for contrast mining, which consists of the discovery of discrepancies between datasets. Lately, it is important to mention the work by Cieslak and Chawla [9], which presents a statistical framework to analyze changes in data distribution resulting in fractures between the data.

A different approach to fixing data fractures relies on the adaptation of the classifier. Quiñonero-Candela et al. [47] edited a very interesting book on the topic, including several specific proposals to repair fractures between data (what they call dataset shift). There are two main differences between the usual proposals in the literature and this contribution: first, they are most often based on altering the classifier, while we propose keeping it intact and transforming the data. Second, most authors focus on covariate shift, a specific kind of data fracture, but the method we propose here is more general and can tackle any kind of shift.

3. Case study: prostate cancer diagnosis

This section begins with an introduction to the importance of the problem in Subsection 3.1. The diagnostic procedure is summarized in Subsection 3.2, and the reason to apply GP-RFD to this problem is shown in Subsection 3.3. Finally, the preprocessing the data went through is presented in Subsection 3.4.

3.1. Motivation

Prostate cancer is the most common non-skin malignancy in the western world. The American Cancer Society estimated 192,280 new cases and 27,360 deaths related to prostate cancer in 2009 [2]. Recognizing the public health implications of this disease, men are actively screened through digital rectal examinations and/or serum prostate specific antigen (PSA) level testing. If these screening tests are suspicious, prostate tissue is extracted, or biopsied, from the patient and examined for structural alterations. Due to imperfect screening technologies and repeated examinations, it is estimated that more than one million people undergo biopsies in the US alone.

3.2. Diagnostic procedure

Biopsy, followed by manual examination under a microscope is the primary means to definitively diagnose prostate cancer as well as most internal cancers in the human body. Pathologists are trained to recognize patterns of disease in the architecture of tissue, local structural morphology and alterations in cell size and shape. Specific patterns of specific cell types distinguish cancerous and non-cancerous tissues. Hence, the primary task of the pathologist examining tissue for cancer is to locate foci of the cell of interest and examine them for alterations indicative of disease. A detailed explanation of the procedure is beyond the scope of this paper and can be found elsewhere [36,41,40].

Operator fatigue is well-documented and guidelines limit the workload and rate of examination of samples by a single operator (examination speed and throughput). Importantly, inter- and intra-pathologist variation complicates decision making. For this reason, it would be extremely interesting to have an accurate automatic classifier to help reduce the load on the pathologists. This was partially achieved in [41], but some issues remain open.

3.3. The generalization problem

Llorà et al. [41] successfully applied a genetics-based approach to the development of a classifier that obtained human-competitive results based on FTIR data. However, the classifier built from the data obtained from one laboratory proved remarkably inaccurate when applied to classify data from a different hospital. Since all the experimental procedure was identical; using the same machine, measuring and post-processing; and having the exact same lab protocols, both for tissue extraction and staining; there was no factor that could explain this discrepancy.

What we attempt to do with this work is develop an algorithm that can evolve a transformation over the data from the second laboratory, creating a new dataset where the classifier built from the first lab is as accurate as possible. This evolved transformation would also provide valuable information, since it would allow the scientists processing the tissues analyze the differences between their results and those of other hospitals.

3.4. Pre-processing of the data

The biological data obtained from the laboratories has an enormous size (in the range of 14 GB of storage per sample); and parallel computing was needed to achieve better-than-human results. For this reason, feature selection was performed on the dataset obtained by FTIR. It was done by applying an evaluation of pairwise error and incremental increase in classification accuracy for every class, resulting in a subset of 93 attributes. This reduced dataset provided enough information for classifier performance to be rather satisfactory: a simple C4.5 classifier achieved ∼95% accuracy on the data from the first lab, but only ∼80% on the second one. The dataset consists of 789 samples from one laboratory and 665 from the other one. These samples represent 0.01% of the total data available for each data set, which were selected applying stratified sampling without replacement. A detailed description of the data pre-processing procedure can be found in [15].

4. A proposal for GP-based feature extraction for the repairing of fractures between data (GP-RFD)

This section is presented in the following way: first, a justification for the choice of GP is included. Subsection 4.1 details how the solutions are represented, then the fitness evaluation procedure and the genetic operators are introduced in Subsections 4.2 and 4.3 respectively. Then, the parameter choices are explained in Subsection 4.4, while the function set is in Subsection 4.5. Finally, the execution flow of the whole procedure is shown in Subsection 4.6.

The problem we are attempting to solve is the design of a method that can create a transformation from a dataset (dataset B) where a classification model is not accurate enough into a new one where it is (dataset S). Said classifier is kept unchanged throughout the process.

We decided to use GP to solve the problem for a number of reasons: first, it is well suited to evolve arbitrary expressions because its chromosomes are trees. This is useful in our case because we want to have the maximum possible flexibility in terms of the functional expressions that can be present in the feature extraction procedure. Second, GP provides highly-interpretable solutions. This is an advantage because our goal is not only to have a new dataset where the classifier works, but also to analyze what was the problem in the first dataset.

The specific decisions to be made once GP was chosen as the technique to apply are how to represent the solutions, what terminals and operators to choose, how to calculate the fitness of an individual and which evolutionary parameters (population size, number of generations, selection and mutation rates, etc.) are appropriate for each specific problem. To clarify some of the points, let us have a binary 2-dimensional problem as an example, and let us use a function set composed of $\{+,-,*,\div\}$.

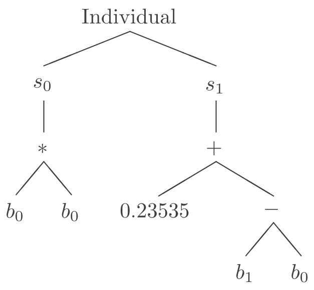
4.1. Solutions representation: context-free grammar

The representation issue was solved by extending GP to evolve more than one tree per solution. Each individual is composed by n trees, where $n = n_v$, the number of attributes present in the dataset (we are trying to develop a new dataset with the same number of attributes as the old one). In the tree structure, the leaves are either constants (we use the Ephemeral Random Constant approach) or attributes from the original dataset. The intermediate nodes are functions from the function set, which is specific to each problem.

The attributes on the transformed dataset are represented by algebraic expressions. These expressions are generated according to the rules of a context-free grammar which allows the absence of some of the functions or terminals. The grammar corresponding to the example problem would look like this:

$$\begin{aligned}
&Start \rightarrow Tree\ Tree\\
&Tree \rightarrow Node\\
&Node \rightarrow Node\ Operator\ Node\\
&Node \rightarrow Terminal\\
&Operator \rightarrow +|-|*|\div\\
&Terminal \rightarrow x_0|x_1|E\\
&E \rightarrow realNumber(represented\ by\ e)
\end{aligned}$$

An individual in the example problem would have two trees; and each of them would be allowed to have any of the functions in the function set, which for this example is $\{+,-,*,\div\}$, in their intermediate nodes; and any of $\{x_0,x_1,e\}$ in the leaves. This, for example, would be a legal individual:



4.2. Fitness evaluation

The fitness evaluation procedure is probably the most treated aspect of design in the literature when dealing with GP-based feature extraction. As has been stated before, the idea is to have the provided classifier's performance drive the evolution. To achieve that, GP-RFD calculates fitness in the following way:

1. Prerequisite: a previously built classifier (the one built from dataset A) needs to be provided. It is used as a black box.
2. Given an individual composed of a list of expression trees (one corresponding to each extracted attribute), a new dataset (dataset S) is built applying the transformations encoded in those expression trees to all the samples in dataset B.
3. The fitness of the individual is the classifier's accuracy on dataset S (training-set accuracy), calculated as the ratio of correctly classified samples over the total number of samples.

Fig. 1 presents a schematic representation of the procedure.

4.3. Genetic operators

This section details the choices made for selection, crossover and mutation operators. Since the objective of this work is not to squeeze the maximum possible performance from GP, but rather to show that it is an appropriate technique for the problem and that it can indeed solve it, we did not pay special attention to these choices, and picked the most common ones in the specialized literature.

- Tournament selection without replacement. To perform this selection, k individuals are first randomly picked from the population (where k is the tournament size), while avoiding using any member of the population more than once. The selected individual is then chosen as the one with the best fitness among those picked in the first stage.
- One-point crossover: for each dimension, a subtree from one of the parents is substituted by one from the other parent. The procedure is specified in Algorithm 1. An example, for one of the dimensions only, can be seen in Fig. 2.

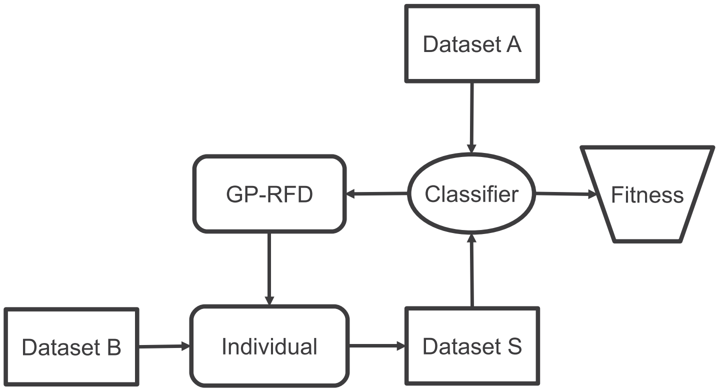


Fig. 1. Schematic representation of the fitness evaluation procedure.

Please cite this article in press as: J.G. Moreno-Torres et al., Repairing fractures between data using genetic programming-based feature extraction: A case study in cancer diagnosis, Inform. Sci. (2010), doi:10.1016/j.ins.2010.09.018

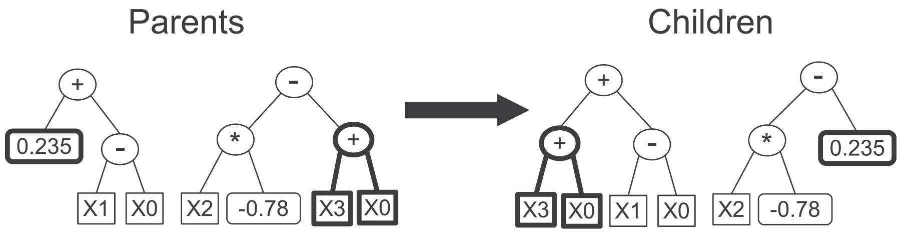


Fig. 2. Crossover example for one of the dimensions only, this is repeated for all dimensions (trees) on each individual.

- Swap mutation: this is a conservative mutation operator, that helps diversify the search within a close neighborhood of a given solution. It consists of exchanging the primitive associated to a node by one that has the same number of arguments.

Algorithm 1. One-point crossover procedure

```
FORALL trees on each individual
1. Randomly select a non-root non-leave node on each of the two parents.
2. The first child is the result of swapping the subtree below the selected node in the father for that of the mother.
3. The second child is the result of swapping the subtree below the selected node in the mother for that of the father.
```

- Replacement mutation: this is a more aggressive mutation operator that leads to diversification in a larger neighborhood. The procedure to perform this mutation is the following:
  1. Randomly select a non-root non-leave node on the tree to mutate.
  2. Create a random tree of depth no more than a fixed maximum depth. This parameter has not been tinkered with, since the goal of this study was not to squeeze the maximum performance out of the proposed method, but rather to check its viability. Future work could tackle this issue.
  3. Swap the subtree below the selected node for the randomly generated one.

4.4. Parameters

The evolutionary parameters that were used for the experimental study are detailed in Table 1. As it was mentioned before, not much attention was payed to optimizing the parameters. Because of this the crossover and mutation probabilities, along with the number of generations to run, were fixed to the usual values in the literature (we could call them 'default values') and were not changed in any of the experiments.

Some of the evolutionary parameters are problem dependent, to select an appropriate value for them we used the following rules:

- Population size: since the only measure of difficulty we know about each of our problems a priori is the number of attributes present in the dataset (n_v), we have to fix the population size as a function of it. In the experiments carried out in this study, we found $400{*}n_v$ to be a large enough population to achieve satisfactory results. This parameter is problem-dependent, so what we are fixing here is an upper bound for the population size needed. We found that, by following this

Table 1
Evolutionary parameters for a n_v-dimensional problem.

Parameter	Value
Number of trees	n_v
Population size	$400{*}n_v$
Duration of the run	50 generations
Selection operator	Tournament without replacement
Tournament size	$log_2(n_v) + 1$
Crossover operator	One-point crossover
Crossover probability	0.9
Mutation operator	Replacement & Swap mutations
Replacement mutation probability	0.001
Swap mutation probability	0.01
Maximum depth of the swapped in subtree	5
Function set	Problem dependent
Terminal set	$\{x_0, x_1, \ldots, x_{n_v} - 1, \mathrm{e}\}$

rule, GP-RFD consistently achieved good results; being able to solve the harder transformations, even though it was excessive for the easier ones and thus resulted in slower execution times. If harder problems than the ones studied in this paper were to be tackled, this parameter might need to be revised.

- Tournament size: since we are increasing the population size as a function of n_v, an increase of the selection pressure is needed too. The formula we used to calculate tournament size is: $log_2(n_v) + 1$. Again, this empirical estimation produced the best results; while an excessive pressure produced too fast of a convergence into local optima, and not enough pressure prevents GP-RFD from converging at all.

Table 2
Datasets used.

Dataset	Attributes	Samples	Classes	Class distribution	Attr. type
Linear synthetic	2	1000	2	50–50%	Real
Tao	2	1888	2	50–50%	Real
Iris	4	150	3	33–33–33%	Real
Phoneme	5	5404	2	70–30%	Real
Wisconsin	9	683	2	65–35%	Real
Heart	13	270	2	55–45%	Real
Wine	13	178	3	33–39%–27%	Real
Wdbc	30	569	2	65–45%	Real
Ionosphere	34	351	2	65–45%	Real
Sonar	60	208	2	54–46%	Real
Mux-11	11	2048	2	50–50%	Nominal
Cancer (A)	93	789	2	60–40%	Real
Cancer (B)	93	665	2	60–40%	Real

Table 3
Transformations performed on the Tao dataset.

Experiment	Rotation	Translate & extrude
Transformation applied	$b_0 = a_0 * cos(1) + a_1 * sin(1)$ $b_1 = a_0 * sin(1) + a_1 * cos(1)$	$b_0 = a_0 * 3 + 2$

Table 4
Transformations performed on the UCI and ELENA datasets.

Dataset	In-set transformation	Out-of-set transformation
Iris	$b_2 = a_2 + a_2$	$b_3 = e^{a_3}$
Phoneme	$b_0 = a_0 - 0.4$ $b_3 = a_3 * 2.5$	$b_0 = sin(a_0)$ $b_3 = cos(a_3)$
Wisconsin	$b_1 = a_1 + 2$ $b_5 = a_5 * 3$	$b_1 = cos(a_1)$ $b_5 = sin(a_5)$
Heart	$b_2 = a_2 * 2$ $b_1 1 = a_1 1 + 3$	$b_2 = sin(a_2)$ $b_1 1 = e^{a_1 1}$
Wine	$b_9 = a_9 - 1$ $b_1 2 = a_1 2 * 2$	$b_9 = sin(a_9)$ $b_1 2 = cos(a_1 2)$
Wdbc	$b_2 6 = a_2 6 - 1$ $b_2 7 = a_2 7 * 3$	$b_2 6 = sin(a_2 6)$ $b_2 7 = cos(a_2 7)$
Ionosphere	$b_4 = a_4 - 0.5$ $b_7 = a_7 * 2$	$b_4 = e^{a_4}$ $b_7 = sin(a_7)$
Sonar	$b_7 = a_7 + 0.3$ $b_4 3 = a_4 3 * 2$	$b_7 = sin(a_7)$ $b_4 3 = e^{a_4 3}$

Table 5
Transformations performed on the Multiplexer-11 dataset.

Experiment	Bit flip	Column swap
Transformation applied	$b_1 = not(a_1)$	$b_1 = a_2$ $b_2 = a_3$ $b_3 = a_1$

Table 6
Experimental parameters.

Dataset	Population size	Tournament size	Function set
Linear synthetic	800	2	$\{+, -, *, \div\}$
Tao	800	2	$\{+, -, *, \div\}$
Iris	1600	3	$\{+, -, *, \div\}$
Phoneme	2000	3	$\{+, -, *, \div\}$
Wisconsin	3600	4	$\{+, -, *, \div\}$
Heart	5200	4	$\{+, -, *, \div\}$
Wine	5200	4	$\{+, -, *, \div\}$
Wdbc	12,000	5	$\{+, -, *, \div\}$
Ionosphere	13,600	6	$\{+, -, *, \div\}$
Sonar	24,000	6	$\{+, -, *, \div\}$
Mux-11	4400	4	$\{+, -, *, \div\}$
Cancer	37,200	6	$\{+, -, *, \div, exp, cos\}$

4.5. Function set

Which functions to include in the function set are usually dependent on the problem. , , , Since one of our goals is to have an algorithm as universal and robust as possible, where the user does not need to fine-tune any parameters to achieve good performance; we decided not to study the effect of different function set choices. The used function sets are chosen to be close to the default ones most authors use in the literature, and were extracted in all cases from $\{+, -, *, \div, exp, cos\}$. The benchmark experiments did not use $\{exp, cos\}$, since we intended to test the capability of the method to unveil transformations that did not include functions in the function set. The specific choices for each of the experiments can be seen in Table 6.

4.6. Execution flow

Algorithm 2 contains a summary of the execution flow of the GP procedure, which follows a classical evolutionary scheme. It stops after a user-defined number of generations, providing as a result the best individual (i.e., transformation) it has ever found.

Algorithm 2. Execution flow of the GP procedure

```
1. Randomly create the initial population by applying the context-free grammar presented in Subsection 4.1.
2. Repeat Ng times (where $Ng$ is the number of generations)
    2.1 Evaluate the current population, using the procedure shown in Subsection 4.2.
    2.2 Apply selection and crossover to create a new population that will replace the old one.
    2.3 Apply the mutation operators to the new population.
3. Return the best individual ever seen.
```

5. Experimental study

This section is organized in the following way: to begin with, a general description of the experimental procedure is presented in Subsection 5.1, along with the datasets that we have used for our testing (both the benchmark problems and the prostate cancer dataset); and also in the benchmarks' case the transformations performed on each of them. The parameters used for each experiment are shown in Subsection 5.2; followed by a presentation of the benchmark experimental results in Subsection 5.3. Finally, the results obtained on the prostate cancer problem are presented in Subsection 5.4.

5.1. Experimental framework, datasets and transformations

The goal of the experiments was to check how effective GP-RFD was in finding a transformation over dataset B that would increase the provided classifier's accuracy. To validate our results, we employed a 5-fold cross validation technique [29]. We used the beagle library [17] for our GP implementation.

The experimental study is fractioned in two parts. In the first one, a synthetic set of tests is built from a few well-known benchmark datasets. The procedure followed in these experiments was (see Fig. 3 for a schematic representation):

1. Split the original dataset in two halves with equal class distribution.
2. Consider the first half, to be dataset A.
3. From dataset A, build a classifier. We chose C4.5 [48], but any other classifier would work exactly the same; due to the fact that GP-RFD uses the learned classifier as a black box.

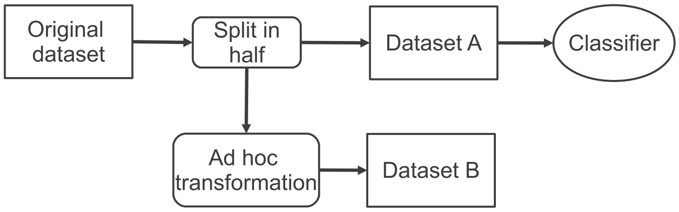


Fig. 3. Schematic representation of the experimental procedure with benchmark datasets.

4. Apply a transformation over the second half of the original dataset, creating dataset B. The transformations we tested were designed to check GP-RFD's performance on different types of problems, including both linear and non-linear transformations. A description of each of them can be found in the next subsection.
5. The performance of the classifier built in step 2 is significantly worse on dataset B than it is on dataset A. This is the starting point on the real problem we are emulating.
6. Apply GP-RFD to dataset B in order to evolve a transformation that will create a solution dataset S. Use 5-fold cross validation over dataset S, so that training and test set accuracy results can be obtained.
7. Check the performance of the step 2 classifier on dataset S. Ideally, it should be close to the one on dataset A, which would mean GP-RFD has successfully discovered the hidden transformation and inverted.

The second part of the study is the application of the proposed algorithm to the prostate cancer problem. The steps followed in this case were:

1. Consider each of the provided datasets to be datasets A and B respectively.
2. From dataset A, build a classifier. Use 5-fold cross validation to obtain training and test-set performance results.
3. Apply GP-RFD to dataset B in order to evolve a transformation that will create a solution dataset S. Use 5-fold cross validation over dataset S, so that training and test set accuracy results can be obtained.
4. Check the performance of the step 2 classifier on dataset S. Ideally, it should be close to the one on dataset A, meaning GP-RFD has successfully discovered the hidden transformation and inverted it.

The selected datasets are summarized in Table 2. A short description and motivation for each of the datasets follows, and this subsection is concluded with the specification of the transformations that were fabricated to test the algorithm on each of the benchmark datasets. For the two-dimensional problems, the transformations are also graphically represented.

Note that the transformations in the prostate cancer problem are not specified. This is due to it being a real-world problem and not a fabricated one, so the implicit transformations in the data were unknown a priori.

- Linear synthetic dataset: we have called the first dataset 'Linear synthetic'. It was created specifically for this work, with the idea of having an easily representable linearly separable dataset to work with. It was chosen to check the performance of GP-RFD on some simple transformations, without the added difficulty of having a complex original dataset. The dataset can be seen in Fig. 4. We applied three transformations to this dataset A: rotation, translation and extrusion and circle. The transformed datasets (datasets B on the experiments) can be seen in Figs. 5–7 respectively.

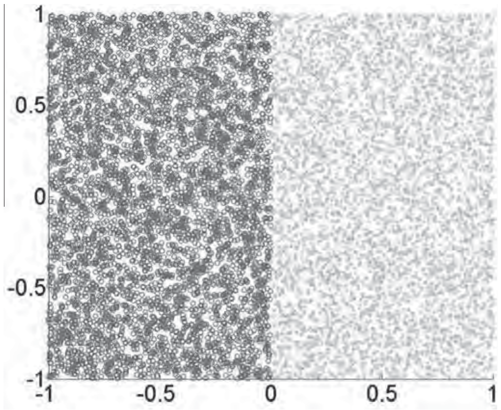


Fig. 4. Linear synthetic dataset, dataset A.

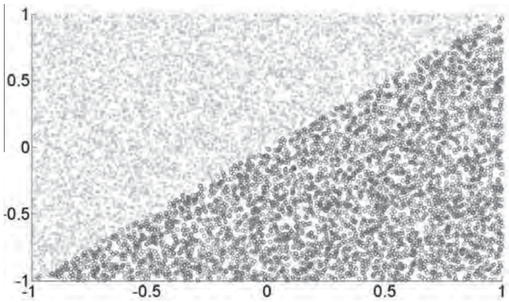


Fig. 5. Rotation problem, transformed dataset.

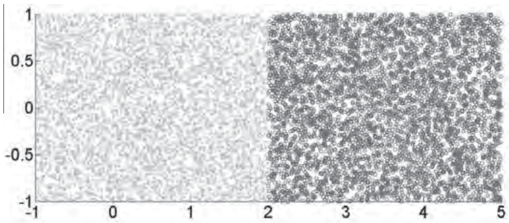


Fig. 6. Translation & extrusion problem, transformed dataset.

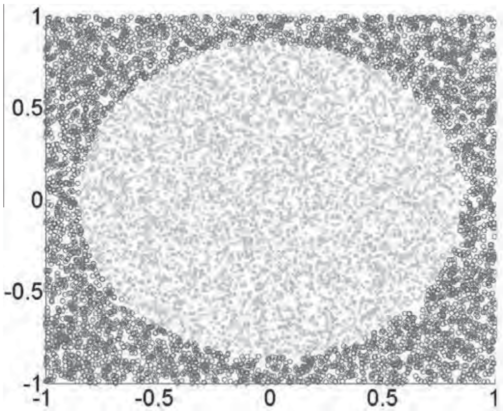


Fig. 7. Circle problem, transformed dataset.

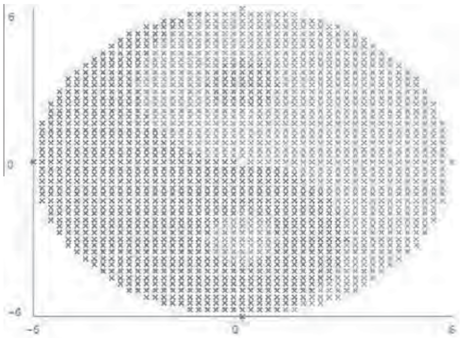


Fig. 8. Tao dataset. This is dataset A, over which the different transformations are applied, and the transformed datasets have to fit to the same classifier this dataset does.

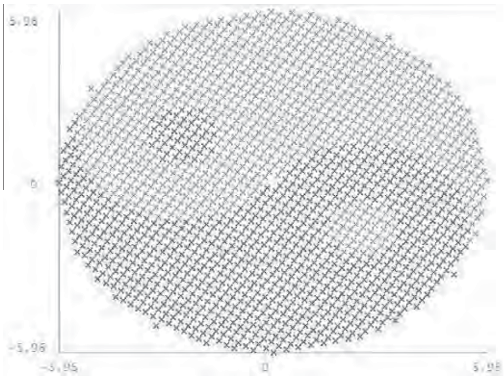


Fig. 9. Rotated Tao, transformed dataset.

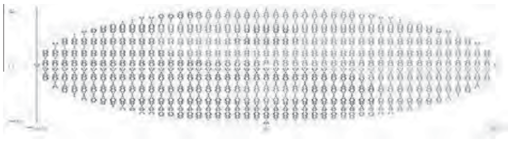


Fig. 10. Translated and extruded Tao, transformed dataset.

- Tao: the next step to check the usefulness of GP-RFD is starting from a harder dataset. To this end, we chose the Tao dataset, still a 2-dimensional problem but where classification is much harder. This dataset is also built artificially [39]. The dataset can be seen, before any transformations (dataset A), in Fig. 8. Mirroring the transformations applied over the linear synthetic dataset, we chose to transform the original Tao dataset by rotating it (Fig. 9); or by translating and extruding (Fig. 10). The transformations applied to Tao can also be seen in Table 3.
- UCI and ELENA datasets: once GP-RFD has been tested in small (with a low number of attributes) datasets, it is useful to see how it fares in bigger benchmark problems. We chose a few different datasets from the UCI database [4], as well as the ELENA project [22]:
  - Iris: classification of iris plants (UCI).
  - Phoneme: distinguish between nasal and oral sounds (ELENA).
  - Wisconsin: diagnosis of breast cancer patients (UCI).
  - Heart: detect the absence or presence of heart disease (UCI).
  - Wine: classification of different types of Italian wines (UCI).
  - Wdbc: determination of whether a found tumor is benign or malignant (UCI).
  - Ionosphere: radar data where the task is to decide is a given radar return is good or bad (UCI, modified as found in the KEEL database [1]).
  - Sonar: distinguishing between rocks and metal cylinders from sonar data (UCI).

We performed two different experiments on each of the datasets. In the first experiment, the transformation is created using functions that appear in the function set of the GP procedure (more specifically, one of the attributes is added to itself). We named this experiment 'in-set transformation'. The second one transforms the dataset by using functions that do not appear in the GP function set. The name for this experiment is 'out-of-set transformation'. The exact details for these transformations can be found in Table 4. Any attribute not specified as being part of the transformation in the tables is assumed to be unchanged.

- Multiplexer-11: since GP-RFD should be flexible enough to be able to tackle datasets with nominal attributes, one of these datasets was included in the testing. In this work, we chose the Multiplexer problem. This is a binary problem where some of the bits act as address, and the remaining bits are data registers. The correct classification for a given input is the value of the register pointed by the address bits. The specific instance used here is Multiplexer-11, a dataset with 11 binary attributes (where the first three act as address, and the remaining eight as registers); and $2^{11} = 2048$ samples. Two different transformations were tested: in the first one, of the address bits was flipped; while in the second experiment there was an attribute swap, in a circular shift. The details can be found in Table 5.
- Prostate cancer: as was explained in Section 3, the solution to this problem is the main motivation for this work. Since we were provided with data from two real laboratories, there was no need to fabricate any transformations: we chose one the data from one of the laboratories as dataset A and the other one as dataset B.

5.2. Parameters

In this section, we detail the parameters used for each of the datasets, including both the evolutionary parameters and the GP setup. The parameters were chosen following the rules detailed in Section 4.4.

As can be seen in Table 6, the population sizes are large. This is mostly due to GP being a technique that traditionally requires large population sizes to be effective, a factor which is aggravated by the fact that GP-RFD evolves multiple expression trees simultaneously (one for each attribute in the dataset). We acknowledge this issue provokes long execution times for some of the experiments, but considered it a secondary concern and did not address it in this work.

5.3. Experimental results: benchmark problems

This part presents the results obtained in terms of classifier performance for the benchmark problems, along with a statistical analysis to evaluate whether GP-RFD is effective.

Table 7 details the performance obtained by the C4.5 classifier on each of the benchmark problems. It includes the classifier performance, calculated as shown on Subsection 4.2, on:

- Dataset A, which was used to generate the decision tree. A 5-fold cross validation technique was applied, and both training and test set results are presented.
- Dataset B, which was created by designing an ad hoc transformation.
- Dataset S, which is the result of applying GP-RFD to dataset B, obtaining a transformed dataset where classifier performance is increased. A 5-fold cross validation technique was applied, and both training and test set results are presented.

The results show that GP-RFD is capable of reversing nearly all of the fabricated transformations, achieving accuracy rates that are very close to the ones obtained in the original datasets in both training and test performances. GP-RFD has also proven capable of generalizing well, as can be seen by the small difference between training and test set classification performances in most cases. However, some of the datasets (which, coincidentally, tend to also behave badly in terms of generalization when building classifiers) present some generalization issues, leading to the inability to fully solve the problem dataset.

5.3.1. Statistical analysis

To complete the experimental study, we have performed a statistical comparison between the classifier performance over the following datasets:

- Dataset A, from which the classifier was built.
- Dataset B, artificially built by injecting an ad hoc transformation.

Table 7
Classifier performance results: benchmark problems.

Problem	Classifier performance on dataset …				
	A-training	A-test	B	S-training	S-test
Linear synthetic – rotation	1.00000	1.00000	0.24930	1.00000	1.00000
Linear synthetic – translation& extrusion	1.00000	1.00000	0.34160	1.00000	0.99800
Linear synthetic – circle	1.00000	1.00000	0.49860	0.96050	0.94400
Tao – rotation	0.98518	0.93750	0.62924	0.94418	0.94255
Tao – translation& extrusion	0.98518	0.93750	0.80403	0.95344	0.93192
Iris – in-set functions	0.97330	0.93333	0.66667	0.99333	0.92000
Iris – out-of-set functions	0.97330	0.93333	0.60000	0.99000	0.92000
Phoneme – in-set functions	0.91895	0.84160	0.75204	0.828978	0.769907
Phoneme – out-of-set functions	0.91895	0.84160	0.59141	0.839871	0.804815
Wisconsin – in-set functions	0.97361	0.93842	0.35380	0.98248	0.93821
Wisconsin – out-of-set functions	0.97361	0.93842	0.88889	0.98321	0.94412
Heart – in-set functions	0.89630	0.72593	0.45296	0.92778	0.79259
Heart – out-of-set functions	0.89630	0.72593	0.60000	0.96296	0.72594
Wine – in-set functions	0.97727	0.89733	0.65556	0.98889	0.90000
Wine – out-of-set functions	0.97727	0.89733	0.40000	0.96944	0.91111
Wdbc – in-set functions	0.98571	0.92143	0.57143	0.98839	0.946428
Wdbc – out-of-set functions	0.98571	0.92143	0.82857	0.98214	0.97500
Ionosphere – in-set functions	0.98286	0.87429	0.70857	0.98571	0.88571
Ionosphere – out-of-set functions	0.98286	0.87429	0.77714	0.98571	0.857143
Sonar – in-set functions	0.93939	0.60601	0.61000	0.95500	0.66000
Sonar – out-of-set functions	0.93939	0.60601	0.51000	0.94750	0.72000
Mux11 – bit flip	1.00000	0.97070	0.50000	0.96951	0.96667
Mux11 – column swap	1.00000	0.97070	0.62500	0.97195	0.96765

- Dataset S-test, the result of applying GP-RFD over dataset B (test-set results).

In [12,18–20] a set of simple, safe and robust non-parametric tests for statistical comparisons of classifiers are recommended. One of them is the Wilcoxon signed-ranks test [55,51], which is the test that we have selected to do the comparison.

This is analogous to the paired t-test in non-parametric statistical procedures; therefore it is a pairwise test that aims to detect significant differences between two sample means, that is, the behavior of two algorithms. It is defined as follows: let d_i be the difference between the performance scores of the two classifiers on the *ith* dataset out of N_{ds} datasets. The differences are ranked according to their absolute values; average ranks are assigned in the case of ties. Let R^+ be the sum of ranks for the data-sets in which the first algorithm outperformed the second, and R^- the sum of ranks for the opposite. Ranks of $d_i = 0$ are split evenly among the sums; if there is an odd number of them, one is ignored:

$$\begin{aligned} R^+ &= \sum_{d_i>0} ank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i) \\ R^- &= \sum_{d_i<0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i) \end{aligned} \tag{1}$$

Let T be the smaller of the sums, $T = min(R^+, R^-)$. If T is less than or equal to the value of the distribution of Wilcoxon for N_{ds} degrees of freedom [59], the null hypothesis of equality of means is rejected; this will mean that a given classifier outperforms their opposite, with the p-value associated.

The Wilcoxon signed-ranks test is more sensitive than the t-test. It assumes commensurability of differences, but only qualitatively: greater differences still count for more, which is probably desired, but the absolute magnitudes are ignored. From a statistical point of view, the test is safer since it does not assume normal distributions. Also, outliers (extremely good/bad performances) have a smaller effect on the Wilcoxon signed-ranks test than on the t-test.

When the assumptions of the paired t-test are met, the Wilcoxon signed-ranks test is less powerful than the paired t-test. On the other hand, when the assumptions are not met, the Wilcoxon test is a better choice than the t-test. This is because the Wilcoxon test can be applied over the averaged results obtained by the algorithms in each data set, without any assumptions about the characteristics of the distribution of the results obtained.

A complete description of the Wilcoxon signed ranks test and other non-parametric tests for pairwise and multiple comparisons, together with software for their use, can be found in the website available at http://sci2s.ugr.es/sicidm/.

As it was mentioned above, the test was applied to compare the classifier performance in datasets A, B and S. The results can be seen in Table 8. Note that we compare the results in dataset A against those in S both in terms of training and test sets. However, since the classifier was not built from dataset B, we consider those results test-set related and compare it with S-test.

So we can conclude GP-RFD is capable of finding transformations resulting in a new dataset S that

1. Significantly outperforms dataset B in terms of classifier performance.
2. Obtains statistically equivalent results to dataset A, both in terms of training and test sets. Since the classifier was built from dataset A, this means dataset S is a successful repair of the fracture between datasets A and B, assuming class

Table 8
Wilcoxon signed-ranks test results: Benchmark problems.

Comparison	R^+	R^-	*p*-Value	Null hypothesis of equality
A-test vs B	275	1	4.77E−007	*rejected* (A-test outperforms B)
B vs S-test	0	276	2.38E−007	*rejected* (S-test outperforms B)
A-training vs S-training	147.5	128.5	–	*accepted*
A-test vs S-test	128.5	147.5	–	*accepted*

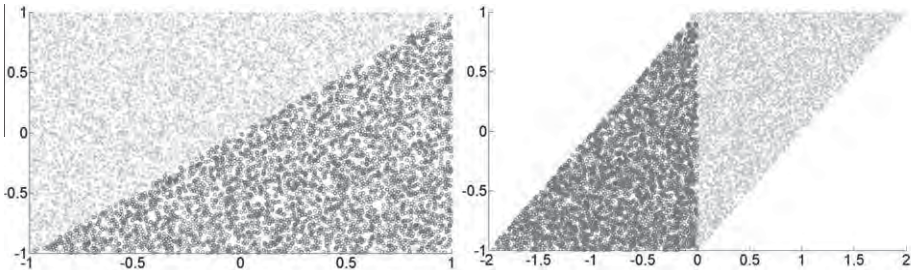


Fig. 11. Linear synthetic rotation, problem (L) and solution (R) datasets.

Please cite this article in press as: J.G. Moreno-Torres et al., Repairing fractures between data using genetic programming-based feature extraction: A case study in cancer diagnosis, Inform. Sci. (2010), doi:10.1016/j.ins.2010.09.018

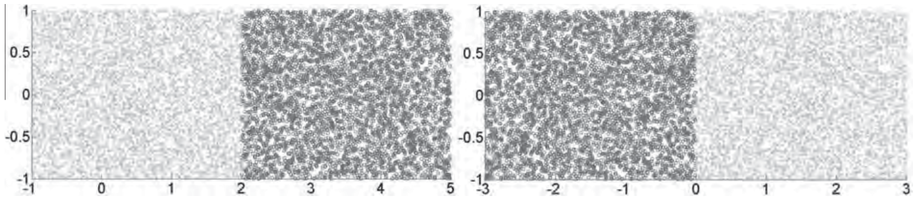


Fig. 12. Linear synthetic translation and extrusion, problem (L) and solution (R) datasets.

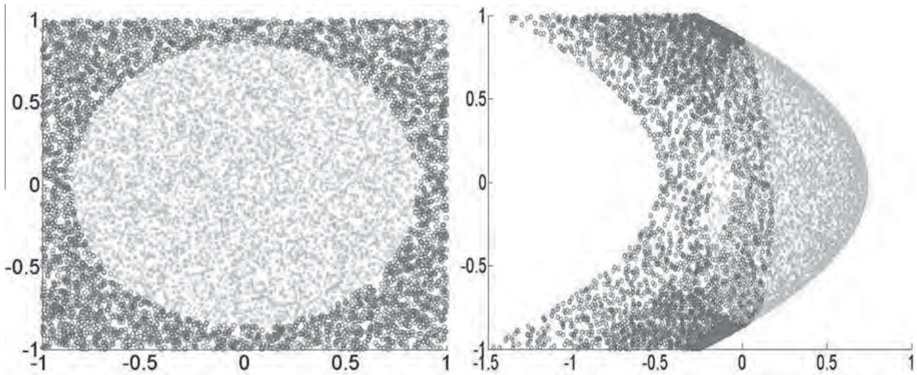


Fig. 13. Circle, problem (L) and solution (R) datasets.

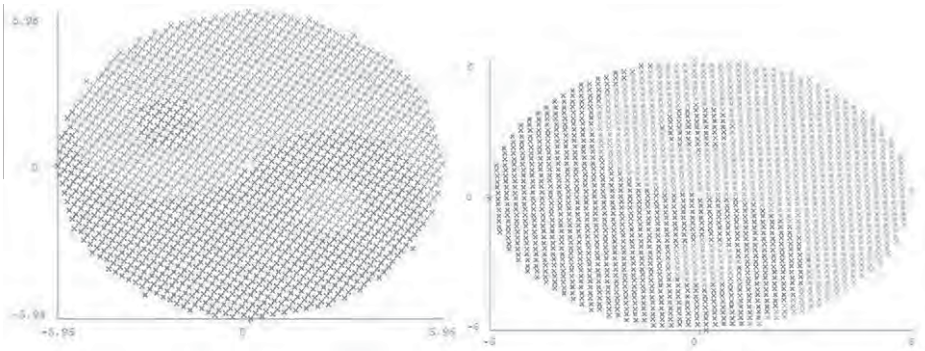


Fig. 14. Rotation in Tao, problem (L) and solution (R) datasets.

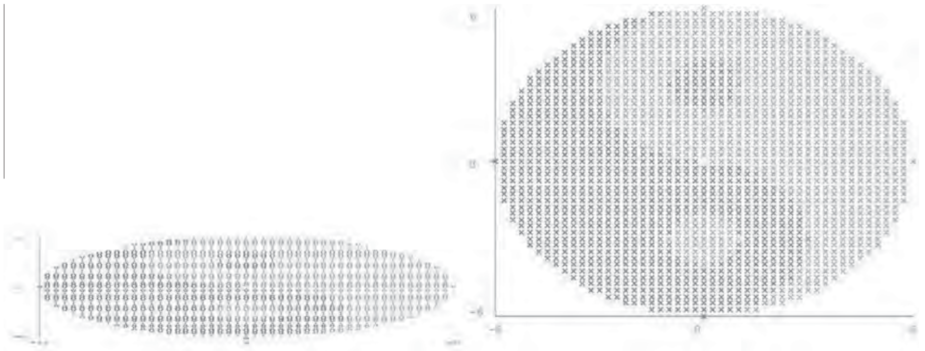


Fig. 15. Translation and extrusion in Tao, problem (L) and solution (R) datasets.

Please cite this article in press as: J.G. Moreno-Torres et al., Repairing fractures between data using genetic programming-based feature extraction: A case study in cancer diagnosis, Inform. Sci. (2010), doi:10.1016/j.ins.2010.09.018

distribution did not change. We know this is the case in these experiments due to the way we built datasets A and B, but it has to be kept in mind when applying the method in other environments.

5.3.2. Graphical results

This section presents graphical representations of some of the obtained results. Since several of the datasets have a high number of variables that make them extremely hard to chart in a simple way, only the results corresponding to the linear synthetic dataset (Figs. 11–13) and the Tao dataset (Figs. 14 and 15) are shown. To make the visualization easier, each of the solution datasets (datasets S) is presented side-by-side with the corresponding problem dataset (datasets B). The original datasets (datasets A) can be seen in Fig. 4 for the linear synthetic dataset and Fig. 8 for the Tao dataset.

5.4. Prostate cancer experimental results

This section presents the preliminary results for the Prostate Cancer problem, in terms of classifier accuracy. The results obtained can be seen in Table 9. In that table, dataset A is the one from the first lab; which was used to build the classifier, dataset B is the one coming from the second lab, and dataset S is the result of the application of GP-RFD.

To check whether the full dataset B was needed to evolve an effective transformation, we also tested using just half of it to train GP-RFD, and the other half to test (2-fold cross validation). These results are also included in Table 9.

The performance results are excellent for a number of reasons. First and foremost, GP-RFD was able to find a transformation over the data from the second laboratory that made the classifier work just as well as it did on the data from the first lab, effectively finding the hidden perturbations that prevented the classifier from working accurately.

The second positive conclusion to be obtained from the results is the generalization power of GP-RFD. As can be observed from the test results, GP-RFD does not ‘cheat’ by over-learning on the known data, and works well when transforming new, previously unseen, samples.

Third, the results show GP-RFD was capable of obtaining excellent results using just half of the B dataset to train. This result highlights the power of the method to unveil the hidden transformation from a relatively low number of samples.

We also performed a Wilcoxon signed-ranks test to evaluate the performance of GP-RFD over the case of study problem. In order to do it, we used the results from each partition in the 5-fold cross validation procedure. We ran the experiment four times, resulting in $4 * 5 = 20$ performance samples to carry out the statistical test. As we did before, R^+ corresponds to the first algorithm in the comparison winning, and R^- to the second one. Table 10 shows the results.

The results on the case study problem are exactly the same as those achieved in the benchmark problems. We can then conclude GP-RFD was capable of repairing the existing fracture between the data from both laboratories. Again, this conclusion assumes class distribution did not change. It is a given in this case, since we know the class distribution to be equal in datasets A and B, but is an issue that has to be kept in mind when applying the method to other problems.

6. Concluding remarks

We have presented GP-RFD, a new algorithm that approaches a common problem in real life for which not many solutions have been proposed in evolutionary computing. The problem in question is the repairing of fractures between data by adjusting the data itself, not the classifiers built from it.

We have developed a solution to the problem by means of a GP-based algorithm that performs feature extraction on the problem dataset driven by the accuracy of the previously built classifier.

Table 9
Classifier performance results: the prostate cancer problem.

Validation method	Classifier performance in dataset ...				
	A-training	A-test	B	S-training	S-test
5-fold cross validation	0.95435	0.92015	0.83570	0.95191	0.92866
2-fold cross validation	0.95435	0.92015	0.83570	0.95482	0.93223

Table 10
Wilcoxon signed-ranks test results: the prostate cancer problem.

Comparison	R^+	R^-	p-Value	Null hypothesis of equality
A-test vs B	210	0	1.91E−007	*rejected* (A-test outperforms B)
B vs S-test	0	210	1.91E−007	*rejected* (S-test outperforms B)
A-training vs S-training	126	84	–	*accepted*
A-test vs S-test	84	126	–	*accepted*

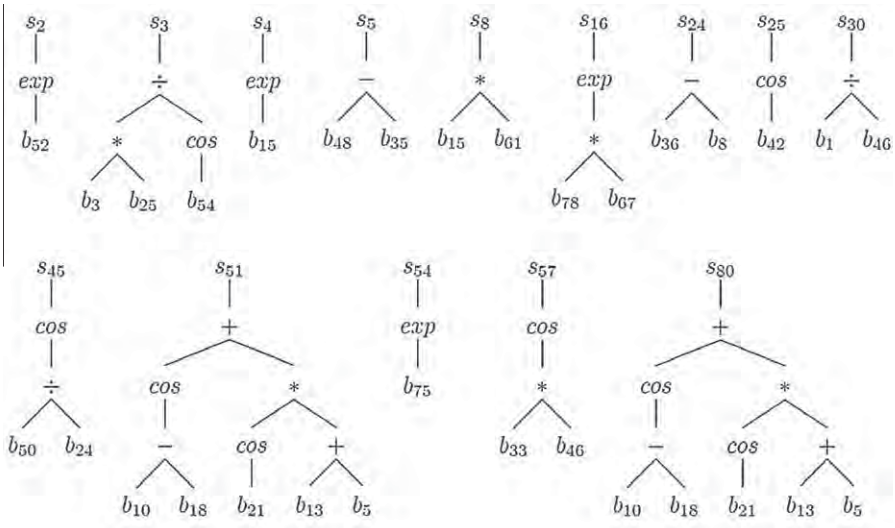


Fig. 16. Tree representation of the expressions contained in a solution to the prostate cancer problem.

We have tested GP-RFD on a set of artificial benchmark problems, where a problem dataset is fabricated by applying an ad hoc disruption to an original dataset, and it has proved capable of solving all the transformations presented showing good performance both in train and, more importantly, test data.

We have also being able to apply GP-RFD to a real-world problem where data from two different laboratories regarding prostate cancer diagnosis was provided, and where the classifier learned from one did not perform well enough on the other. Our algorithm was capable of learning a transformation over the second dataset that made the classifier fit just as well as it did on the first one. The validation results with 5-fold cross validation also support the idea that the algorithm is obtaining good results; and has a strong generalization power.

Lastly, we have applied a statistical analysis methodology that supports the claim that the classifier performance obtained on the solution dataset significantly outperforms the one obtained on the problem dataset.

There is, however, one point where the proposed method has not been successful. The learned transformations have failed to provide any information about why the fracture appeared between the data from the two laboratories. We have, however, included a sample of the transformations learned in appendix A.

Acknowledgments

Jose García Moreno-Torres was supported by a scholarship from 'Obra Social la Caixa' and is currently supported by a FPU grant from the Ministerio de Educación y Ciencia of the Spanish Government, and also by the KEEL project (TIN2008-06681-C06-01). Rohit Bhargava would like to acknowledge collaborators over the years, especially Dr. Stephen M. Hewitt and Dr. Ira W. Levin of the National Institutes of Health, for numerous useful discussions and guidance. Funding for this work was provided in part by University of Illinois Research Board and by the Department of Defense Prostate Cancer Research Program. This work was also funded in part by the National Center for Supercomputing Applications and the University of Illinois, under the auspices of the NCSA/UIUC faculty fellows program.

Appendix A. Sample solution from the prostate cancer problem

In this appendix, we include a sample of the learned transformations for the prostate cancer problem, presenting the transformations corresponding to the highest fitness individual ever found. Due to space concerns, only the attributes relevant to the C4.5 classifier are shown (Fig. 16).

References

[1] J. Alcalá-fdez, L. Sánchez, S. García, M.J.D. Jesus, S. Ventura, J.M. Garrell, J. Otero, J. Bacardit, V.M. Rivas, J.C. Fernández, F. Herrera, Keel: a software tool to assess evolutionary algorithms for data mining problems, Soft Computing - A Fusion of Foundations, Methodologies and Applications 13 (3) (2008) 307–318.

[2] AmericanCancerSociety. How many men get prostate cancer? <http://www.cancer.org/docroot/CRI/content/CRI_2_2_1X_How_many_men_get_prostate_cancer_36.asp>.

[3] A. Arcuri, X. Yao, Co-evolutionary automatic programming for software development, Information Sciences (2010), in press, doi:10.1016/j.ins.2009.12.019.

[4] A. Asuncion, D. Newman, UCI machine learning repository (2007).

[5] W. Banzhaf, F.D. Francone, P. Nordin, The effect of extensive use of the mutation operator on generalization in genetic programming using sparse data sets, in: In Parallel Problem Solving from Nature IV, Proceedings of the International Conference on Evolutionary Computation, Springer Verlag, 1996, pp. 300–309.

[6] F. Berlanga, A. Rivera, M. del Jesus, F. Herrera, GP-COACH: genetic programming-based learning of compact and accurate fuzzy rule-based classification systems for high-dimensional problems, Information Sciences 180 (8) (2010) 1183–1200.

[7] A.S. Bickel, R.W. Bickel, Tree structured rules in genetic algorithms, In Proceedings of the Second International Conference on Genetic Algorithms and their Application, L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 1987.

[8] M.C.J. Bot, Feature extraction for the k-nearest neighbour classifier with genetic programming, In EuroGP '01: Proceedings of the Fourth European Conference on Genetic Programming, Springer-Verlag, London, UK, 2001.

[9] D.A. Cieslak, N.V. Chawla, A framework for monitoring classifiers' performance: when and why failure occurs?, Knowledge and Information Systems 18 (1) (2009) 83–108

[10] N.L. Cramer, A representation for the adaptive generation of simple sequential programs, In Proceedings of the 1st International Conference on Genetic Algorithms, L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 1985.

[11] C. Darwin, On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life, John Murray, London, UK, 1859.

[12] J. Demšar, Statistical comparisons of classifiers over multiple data sets, Journal of Machine Learning Research 7 (2006) 1–30.

[13] P.G. Espejo, S. Ventura, F. Herrera, A survey on the application of genetic programming to classification, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 40 (2) (2010) 121–144.

[14] M. Evett, T. Fernandez, Numeric mutation improves the discovery of numeric constants in genetic programming, in: J. Koza (Ed.), Proceedings of the Third Annual Genetic Programming Conference, Morgan Kaufmann, Madison, WI, 1998, pp. 66–71.

[15] D.C. Fernandez, R. Bhargava, S.M. Hewitt, I.W. Levin, Infrared spectroscopic imaging for histopathologic recognition, Nature Biotechnology 23 (4) (2005) 469–474.

[16] C. Fujiko, J. Dickinson, Using the genetic algorithm to generate lisp source code to solve the prisoner's dilemma, in: Proceedings of the Second International Conference on Genetic Algorithms and their application, L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 1987, pp. 236–240.

[17] C. Gagné, M. Parizeau, Genericity in evolutionary computation software tools: principles and case study, International Journal on Artificial Intelligence Tools 15 (2) (2006) 173–194.

[18] S. García, A. Fernández, J. Luengo, F. Herrera, A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability, Soft Computing 13 (10) (2009) 959–977.

[19] S. García, A. Fernández, J. Luengo, F. Herrera, Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: experimental analysis of power, Information Sciences 180 (10) (2010) 2044–2064.

[20] S. García, F. Herrera, An extension on 'statistical comparisons of classifiers over multiple data sets' for all pairwise comparisons, Journal of Machine Learning Research 9 (2008) 2677–2694.

[21] D.E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley, Reading, MA, 1989.

[22] A. Guérin-Dugué et al. Deliverable R3-B1-P - Task B1: Databases. Technical report, Elena-NervesII "Enhanced Learning for Evolutive Neural Architecture, ESPRIT-Basic Research Project Number 6891, June 1995, Anonymous FTP:/pub/neural-nets/ELENA/Databases.ps.Z on ftp.dice.ucl.ac.be.

[23] H. Guo, A.K. Nandi, Breast cancer diagnosis using genetic programming generated feature, Pattern Recognition 39 (5) (2006) 980–987.

[24] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, Journal of Machine Learning Research 3 (2003) 1157–1182.

[25] I. Guyon, S. Gunn, M. Nikravesh, L. Zadeh (Eds.), Feature Extraction, Foundations and Applications, Springer, 2006.

[26] C. Harris, An investigation into the Application of Genetic Programming techniques to Signal Analysis and Feature Detection, PhD thesis, University College, London, 26 Sept. 1997.

[27] J.H. Holland, Adaptation in Natural and Artificial Systems, University of Michigan Press, Ann Arbor, MI, USA, 1975.

[28] K.-A. Kim, S.-Y. Oh, H.-C. Choi, Facial feature extraction using pca and wavelet multi-resolution images, in: Sixth IEEE International Conference on Automatic Face and Gesture Recognition, IEEE Computer Society, Los Alamitos, CA, USA, 2004, p. 439.

[29] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, Morgan Kaufmann, 1995, pp. 1137–1143.

[30] M. Kotani, S. Ozawa, M. Nakai, K. Akazawa, Emergence of feature extraction function using genetic programming, In KES (1999) 49–152.

[31] P. Kouchakpour, A. Zaknich, T. Bräunl, Population variation in genetic programming, Information Sciences 177 (17) (2007) 3438–3452.

[32] P. Kouchakpour, A. Zaknich, T. Bräunl, Dynamic population variation in genetic programming, Information Sciences 179 (8) (2009) 1078–1091.

[33] J. Koza, Genetic Programming: On the Programming of Computers by Means of Natural Selection, The MIT Press, Cambridge, MA, 1992.

[34] J. Koza, Genetic programming II: Automatic Discovery of Reusable Programs, Complex Adaptive Systems, MIT Press, Cambridge, Mass, 1994.

[35] J.R. Koza, M.J. Streeter, M.A. Keane, Routine high-return human-competitive automated problem-solving by means of genetic programming, Information Sciences 178 (23) (2008) 4434–4452.

[36] I.W. Levin, R. Bhargava, Fourier transform infrared vibrational spectroscopic imaging: integrating microscopy and molecular recognition, Annual Review of Physical Chemistry 56 (2005) 429–474.

[37] J.-Y. Lin, H.-R. Ke, B.-C. Chien, W.-P. Yang, Classifier design with feature selection and feature extraction using layered genetic programming, Expert Systems with Applications 34 (2) (2008) 1384–1393.

[38] H. Liu, H. Motoda, Feature Extraction, Construction and Selection: A Data Mining Perspective, Kluwer Academic, Boston, 1998. vol. SECS 453.

[39] X. Llorà, J.M. Garrell, Knowledge-independent data mining with fine-grained parallel evolutionary algorithms, in: Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2001), Morgan Kaufmann Publishers, 2001, pp. 461–468.

[40] X. Llorà, A. Priya, R. Bhargava, Observer-invariant histopathology using genetics-based machine learning, Natural Computing: An International Journal 8 (1) (2009) 101–120.

[41] X. Llorà, R. Reddy, B. Matesic, R. Bhargava, Towards better than human capability in diagnosing prostate cancer using infrared spectroscopic imaging, in: GECCO '07: Proceedings of the Ninth Annual Conference on Genetic and Evolutionary Computation, ACM, New York, NY, USA, 2007, pp. 2098–2105.

[42] U.-M. O'Reilly, An Analysis of Genetic Programming, PhD thesis, Carleton University, Ottawa-Carleton Institute for Computer Science, Ottawa, Ontario, Canada, 1995.

[43] M. Pei, E.D. Goodman, W.F. Punch. Pattern discovery from data using genetic algorithms, in: Proceeding of First Pacific-Asia Conference Knowledge Discovery & Data Mining(PAKDD-97), 1997.

[44] A. Piszcz, T. Soule, A survey of mutation techniques in genetic programming, in: GECCO '06: Proceedings of the Eigth Annual Conference on Genetic and Evolutionary Computation, ACM, New York, NY, USA, 2006, pp. 951–952.

[45] I.T. Podolak, Facial component extraction and face recognition with support vector machines, in: FGR '02: Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition, IEEE Computer Society, Washington, DC, USA, 2002, p. 83.

[46] R. Poli, W.B. Langdon, N.F. Mcphee, A Field Guide to Genetic Programming, Lulu Enterprises Ltd, UK, 2008.

[47] J. Quiñonero Candela, M. Sugiyama, A. Schwaighofer, N.D. Lawrence, Dataset Shift in Machine Learning, The MIT Press, 2009.

[48] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[49] C. Ryan, M. Keijzer, An analysis of diversity of constants of genetic programming, in: C. Ryan, T. Soule, M. Keijzer, E. Tsang, R. Poli, E. Costa (Eds.), Genetic Programming, Proceedings of EuroGP'2003, LNCS, vol. 2610, Springer-Verlag, Essex, 2003, pp. 404–413.

[50] J.R. Sherrah, R.E. Bogner, A. Bouzerdoum, The evolutionary pre-processor: automatic feature extraction for supervised classification using genetic programming, Proceedings of the Second International Conference on Genetic Programming, vol. GP-97, Morgan Kaufmann, 1997, pp. 304–312.

[51] D.J. Sheskin, Handbook of Parametric and Nonparametric Statistical Procedures, 4th ed., Chapman & Hall/CRC, 2007.

[52] M.G. Smith, L. Bull, Genetic programming with a genetic algorithm for feature construction and selection, Genetic Programming and Evolvable Machines 6 (3) (2005) 265–281.

[53] W.A. Tackett, Genetic programming for feature discovery and image discrimination, in: Proceedings of the Fifth International Conference on Genetic Algorithms, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993, pp. 303–311.

[54] K. Wang, S. Zhou, C.A. Fu, J.X. Yu, F. Jeffrey, X. Yu, Mining changes of classification by correspondence tracing, in: Proceedings of the 2003 SIAM International Conference on Data Mining (SDM 2003), 2003.

[55] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics Bulletin 1 (6) (1945) 80–83.

[56] N. Wyse, R. Dubes, A. Jain, A critical evaluation of intrinsic dimensionality algorithms a critical evaluation of intrinsic dimensionality algorithms, in: E.S. Gelsema, L.N. Kanal (Eds.), Pattern Recognition in Practice, Morgan Kauffman Publishers, Inc., Amsterdam, 1980, pp. 415–425.

[57] Y. Yang, X. Wu, X. Zhu, Conceptual equivalence for contrast mining in classification learning, Data & Knowledge Engineering 67 (3) (2008) 413–429.

[58] A. Zafra, S. Ventura, G3P-MI: a genetic programming algorithm for multiple instance learning, Information Sciences 180 (23) (2010) 4496–4513.

[59] J.H. Zar, Biostatistical Analysis, 5th ed., Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2007.

[60] Y. Zhang, H. Li, M. Niranjan, P. Rockett, Applying cost-sensitive multiobjective genetic programming to feature extraction for spam E-mail filtering, in: Proceedings of the 11th European Conference on Genetic Programming, EuroGP 2008, Lecture Notes in Computer Science, vol. 4971, Springer, Naples, 2008, pp. 325–336.

[61] Y. Zhang, P.I. Rockett, A generic optimal feature extraction method using multiobjective genetic programming, Technical Report VIE 2006/001, Department of Electronic and Electrical Engineering, University of Sheffield, UK, 2006.

[62] Y. Zhang, P.I. Rockett, A generic multi-dimensional feature extraction method using multiobjective genetic programming, Evolutionary Computation 17 (1) (2009) 89–115.

Please cite this article in press as: J.G. Moreno-Torres et al., Repairing fractures between data using genetic programming-based feature extraction: A case study in cancer diagnosis, Inform. Sci. (2010), doi:10.1016/j.ins.2010.09.018

Shining a new light into molecular workings

Francis L Martin

A technique to substantially increase the resolution and imaging area of Fourier-transform infrared microspectroscopy, while decreasing the amount of time required for image acquisition, may augment the use of this technology in biomedical and environmental research.

The application of infrared spectroscopy technologies has gained increasing recognition in recent years as an adjunct support to more traditional methodologies, especially in cell biology[1–3]. A report from Hirschmugl, Bhargava and co-workers, in this issue of *Nature Methods*[4], demonstrates a two orders of magnitude improvement to this technique, which for the first time genuinely allows the acquisition of intracellular chemical information with better than micrometer-scale spatial resolution. Additionally, this group provides the scientific community with an experimental setup for acquisition of minute-by-minute spectral information over the entire mid-infrared region with an excellent signal-to-noise ratio; this could be used to nondestructively monitor living biological material. In the emerging field of biospectroscopy, which has seen several pioneering developments over the last two decades[5,6], this work has the genuine potential to act as a bridge for implementing infrared spectroscopy into mainstream biological practice.

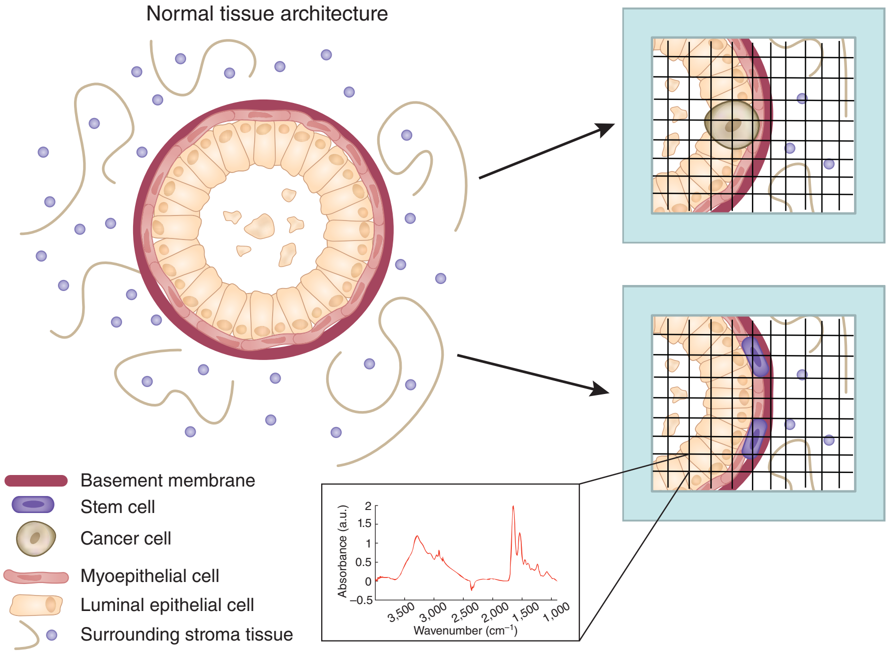


Figure 1 | Chemical imaging of different tissue components in a pixel-by-pixel fashion. The method of Nasse *et al.*[4] could be applied to different applications within normal-looking tissue architecture such as identifying early cancerous cells or the *in situ* location of stem cells required to regenerate a tissue. Fast imaging at high spatial resolution (<1 micrometer) in living tissue would even allow for analyses of cell membranes or organelles. Different components would be identified on the basis of location of their unique spectral signatures.

Fourier-transform infrared (FTIR) spectroscopy for biological, environmental or biomedical applications exploits the fact that biomolecules absorb in the mid-infrared frequency range in a manner consistent with the chemical-bond composition of the interrogated sample. Based on the absorbance pattern of the present chemical bonds with an electric dipole moment that changes during vibration, signature spectra are derived. Applied in imaging format, different FTIR microscopy platforms are available to obtain such chemical information, from a benchtop instrument that can be found in a typical physics or chemistry department to an infrared beamline in one of the 50 or so synchrotron facilities worldwide. A benchtop instrument delivers reasonable spectral information but limited spatial resolution; for instance, it might allow one to derive an integrated tissue spectral signature. A synchrotron system typically yields a higher signal-to-noise ratio and can also approximate single-cell spatial resolution.

Of course, not everyone has ready access to a synchrotron facility (which typically require applications for beam time) and often such facilities might not be developed sufficiently to integrate the requirements of a biological laboratory. The time-consuming nature of individual experiments within a limited beam time allocation currently minimizes the number of replicates achievable, which lessens the robustness of the findings. Consequently, the approach of infrared spectroscopy to biological or environmental questions often appears exotic and niche.

FTIR spectroscopy has long been applied in the physical sciences as it yields chemical

Francis L. Martin is at the Centre for Biophotonics, Lancaster Environment Centre, Lancaster University, Lancaster, UK.
e-mail: f.martin@lancaster.ac.uk

information about a sample. To apply the technique to biological specimens—from sample preparation to understanding the limitations of the system at hand to processing and interpreting the acquired spectra—the practitioner requires a truly interdisciplinary approach[7]. Infrared microspectroscopy data collection has been slow, and the spatial resolution has been poor. In addition, as computer processing capabilities have increased, spectral datasets have grown increasingly large; thus, developing and implementing appropriate computational algorithms capable of extracting relevant biomarkers is yet another cross-discipline hurdle[8]. As we better understand the nature of such derived spectra and the underlying physical phenomena that may modify their structure, our processing of them and as such the information we extract from them will undoubtedly evolve[9]. Within this interdisciplinary milieu is the conundrum as to why infrared microspectroscopy remains so under-exploited in the cell biology arena.

The advantage of applying this technology in an imaging format is that neither labeling nor staining of the sample is required. Whereas with other microscopy methods one needs a priori knowledge of the sample to be interrogated to facilitate the tracking of a prescribed biomolecule, unlimited by this restriction, infrared microspectroscopy is a discovery-based method. However, at the same time, the inability to track specific molecules with infrared microspectroscopy means that it should be used as a complementary tool to optical microscopy, not as a replacement for it.

The approach implemented by Hirschmugl, Bhargava and co-workers[4] that allows spectral imaging at high resolution over a wide surface area (a tissue section containing several glandular elements) in a short time frame (within minutes) may substantially increase the application of FTIR microspectroscopy in biological research, allowing it to provide complementary information to that of optical microscopy techniques. They achieved this by harnessing multiple synchrotron beams to a focal plane array detector; the latter is an imaging device consisting of an array of light-sensing pixels placed at the lens' focal plane.

The fact that FTIR microscopy image data now can be acquired in minutes rather than hours or days in an evolving live cell context opens up new avenues of investigation, from tracking the cell differentiation process[2], to shedding new insights into cell cycle kinetics to nanoparticle-induced toxicity mechanisms, to looking at host-parasite interactions[10]. In addition, given that infrared imaging is nondestructive, this allows one to match the point-by-point spectral information to more conventional staining approaches that can be subsequently applied; thus, one might envision imaging a tissue section containing a glandular element suspended in stroma and then matching the location of spectral signatures of suspect transformed cells or quiescent stem cells to more conventional biomarkers (**Fig. 1**). This study[4] applies the approach to sections of prostate and breast tissue; the resultant images show remarkable clarity at a spatial resolution hitherto unachievable with infrared spectroscopy and over a wider area with a much shorter acquisition time.

This advance now leads the way to stimulating the development of new infrared microscopy systems for routine use. Currently, new light sources, such as quantum cascade lasers, are being developed that it is expected will be capable of generating the requisite brilliance of light in a benchtop system and thus will have the potential to generate a similar level of spectral quality and spatial resolution as a synchrotron-harnessed system, which will be necessary to truly make infrared microspectroscopy a core technology in the biology laboratory.

Additionally, the technique may have important clinical and environmental applications; infrared microspectroscopy will potentially allow a pathologist to examine an image of cellular architecture as well as reference the underlying signature chemical information inherent in this picture. Although traditional methods such as hematoxylin-and-eosin staining are ingrained in clinical practice, such approaches are often fraught with subjectivity and lack molecular detail such as conclusive evidence of the earliest predisease alterations. Infrared approaches allow for objective insight, which could facilitate earlier diagnosis, which would allow for enormous societal benefits. In environmental research, there is an urgent need for new approaches to monitor sentinel organism effects after complex exposures; again, computational algorithms can be exploited to extract such mechanistic information after spectral analyses[11].

Galileo used an optical telescope to peer into the universe. In the twentieth century, astronomers began to use infrared telescopes. Likewise, biologists in a multiplicity of different disciplines have used optical microscopes to understand biological architecture and function. Infrared microscopes shine a new light into the microverse that is the biological cell, allowing one to visualize processes differently or to identify novel components that may facilitate better explanation or understanding. This is not to say that infrared microscopy will one day replace optical microscopy or other conventional methodologies, but it may explain phenomena that hitherto would not have been explained using traditional approaches.

The pioneering work by Hirschmugl, Bhargava and co-workers[4] adds to the impetus toward developing benchtop instruments with similar capability in the biological laboratory.

COMPETING FINANCIAL INTERESTS

The author declares no competing financial interests.

1. Baker, M.J. *et al. Br. J. Cancer* **99**, 1859–1866 (2008).
2. Walsh, M.J. *et al. Stem Cells* **26**, 108–118 (2008).
3. Romeo, M., Mohlenhoff, B., Jennings, M. & Diem, M. *Biochim. Biophys. Acta* **1758**, 915–922 (2006).
4. Nasse, M.J. *et al. Nat. Methods* **8**, 413–416 (2011).
5. Naumann, D., Helm, D. & Labischinski, H. *Nature* **351**, 81–82 (1991).
6. Malins, D.C., Polissar, N.L. & Gunselman, S.J. *Proc. Natl. Acad. Sci. USA* **93**, 14047–14052 (1996).
7. Martin, F.L. *et al. Nat. Protoc.* **5**, 1748–1760 (2010).
8. Kelly, J.G. *et al. J. Proteome Res.* advance online publication. doi:10.1021/pr101067u (6 January 2011).
9. Bassan, P. *et al. Analyst* **134**, 1586–1593 (2009).
10. Heraud, P., Wood, B.R., Tobin, M.J., Beardall, J. & McNaughton, D. *FEMS Microbiol. Lett.* **249**, 219–225 (2005).
11. Llabjani, V., Trevisan, J., Jones, K.C., Shore, R.F. & Martin, F.L. *Environ. Sci. Technol.* **44**, 3992–3998 (2010).

THE AUTHOR FILE

Rohit Bhargava and Carol Hirschmugl

Multiple synchrotron beams make infrared imaging faster and clearer.

Rohit Bhargava expected that 12 beams from a synchrotron would boost the quality of his imaging data, but he still was not prepared for what he saw. Bhargava, a bioengineer at the University of Illinois at Urbana-Champaign, uses a technique called infrared spectroscopic imaging, which detects various chemical groups in a sample based on their absorbance of infrared light. It is a specialized technique, valued not for its ability to produce stunning pictures but because it can yield molecular-level information without the need for labels. Normally, though, that means detecting lipids, fats and carbohydrates with a resolution of 5 micrometers or worse. Hence Bhargava's surprise at results from the new technique: "You could start to see details that you are used to seeing only in optical microscopy," he recalls. "The crispness, the details were comparable." In fact, the pixel size is only a half micrometer in diameter, a hundredth the size of current state-of-the-art infrared imaging.

Carol Hirschmugl and Rohit Bhargava

"Even to this day, every time we take data, I'm shocked by the quality of the images," says Carol Hirschmugl, a physicist at the University of Wisconsin–Milwaukee who, with Michael Nasse and others, developed a technique that uses multiple synchrotron beams to illuminate samples with infrared light. Not only is resolution improved, the imaging technique is also faster than methods that rely on single beams or on heat sources to produce infrared radiation. Data that would normally take 11 days to collect can now be acquired in 20 minutes.

The idea for the project began when Hirschmugl learned about a 'weekend experiment' at Brookhaven National Laboratories. A team of scientists set four beams onto nonbiological samples and showed that resolution improved, she says, but they did not take the project further to complex biological samples, largely because getting access to a beamline is difficult.

Hirschmugl was intrigued, so she approached the scientists at the Synchrotron Radiation Center at the University of Wisconsin–Madison, which was built to produce beams with considerably less noise than other synchrotrons. The director offered access to a bank of 12 beams—provided Hirschmugl could get the necessary funding.

When the funding came in, she and the engineers got access to the beamline within 2 years (a timeline so short Hirschmugl refers to it as "miraculous"). To figure out how to harness the setup for infrared imaging, the team spent weeks, sometimes working "morning to morning," testing algorithms to align the beams and focus the 48 mirrors.

Their technique could be used to take great pictures of polystyrene beads, but to learn whether the method would be useful for biological imaging, Hirschmugl's team had to find a biologist who could ask the right kinds of questions. That led her to Bhargava, who had worked on some of the earliest prototypes of infrared microscopy, including theoretical research on how to acquire data to get informative images.

"Out of the blue I got a call from Carol," Bhargava recalls. She said she had an interesting instrument and invited him to try it out. "It was very clear from the theoretical work that this would be something different," he says, "but I couldn't have anticipated the nice results we would get." For the first time, the researchers could distinguish the collagen-dense interface between epithelial and stromal cells using infrared imaging, and could distinguish between cancerous and healthy tissue in fixed slides of prostate and breast samples.

But that is just the beginning, the collaborators say. Any samples that have chemical organization at the micrometer scale can be imaged in this facility: projects under way include studying stem-cell differentiation, malarial parasites inside cells and even the pigment and oil layers of 500-year-old paintings. Theory-based research can also expand. Work developed for wide-field imaging with optical techniques can be applied to infrared imaging, and experiments on the synchrotron may show ways to improve desktop infrared imaging instruments.

Hirschmugl plans to invite more collaborators to the facility and even to build facilities onsite to enable experiments on living cell cultures. That, however, may depend on new sources of funding: in the same month this paper in *Nature Methods* was accepted for publication, the US National Science Foundation cut funds for the Synchrotron Radiation Center. "Now that we have these beautiful results, [the National Science Foundation] is not funding the running of the synchrotron," says Hirschmugl. "It's been an up and down time." But perhaps, she says, it represents a new opportunity; she and colleagues are looking for funding sources to reinvent the Synchrotron Radiation Center as a dedicated infrared imaging center.

Monya Baker

Nasse, M.J. *et al.* High-resolution Fourier-transform infrared chemical imaging with multiple synchrotron beams. *Nat. Methods* **8**, 413–416 (2011).

High-resolution Fourier-transform infrared chemical imaging with multiple synchrotron beams

Michael J Nasse[1,2], Michael J Walsh[3], Eric C Mattson[1], Ruben Reininger[4], André Kajdacsy-Balla[5], Virgilia Macias[5], Rohit Bhargava[3] & Carol J Hirschmugl[1]

Conventional Fourier-transform infrared (FTIR) microspectroscopic systems are limited by an inevitable trade-off between spatial resolution, acquisition time, signal-to-noise ratio (SNR) and sample coverage. We present an FTIR imaging approach that substantially extends current capabilities by combining multiple synchrotron beams with wide-field detection. This advance allows truly diffraction-limited high-resolution imaging over the entire mid-infrared spectrum with high chemical sensitivity and fast acquisition speed while maintaining high-quality SNR.

Stains and labels to enhance contrast in microscopy have been used for many years, leading to many important discoveries. However, their use is often time-consuming and cumbersome, can perturb the function of drugs or small metabolites or may be cytotoxic. In contrast, label-free chemical imaging requires no artificial modification of biomolecules or additional sample preparation and permits a comprehensive characterization of heterogeneous materials[1]. Chemical imaging is generating considerable interest for biomedical analysis as dyes or stains are not required for contrast and substantial chemical and structural information can be extracted without prior knowledge of molecular epitopes or manual interpretation. Vibrational spectroscopic techniques, including both mid-infrared absorption and Raman scattering–based imaging, permit molecular analyses without perturbation. Spontaneous Raman scattering relies on a very weak effect and therefore involves a trade-off between measurement time and sensitivity, potentially leading to photoinduced sample damage. Emerging instrumentation[2] involving nonlinear Raman contrast has considerably extended imaging capabilities beyond these traditional trade-offs, and exciting work is underway to carefully match lasers and reject spurious backgrounds (for example, in coherent anti-Stokes Raman scattering) and in extending wavelength coverage and speed (for example, in stimulated Raman scattering). Conversely, the strong mid-infrared absorption contrast makes infrared spectroscopy and microscopy a straightforward, nondestructive, label-free chemical contrast modality with broad applications[1,3] ranging from the analysis of graphene-based materials, pharmaceuticals, volcanic rocks and biominerals to applications in forensics and art conservation, among others. Infrared spectroscopic tools are particularly interesting for applications in biomedical fields such as marine biology, cancer research, stem cells (for example, to delineate cell mechanisms or lineage), real-time monitoring of live cells, Alzheimer's disease, Malaria parasites and more[3] (Online Methods).

Infrared instrumentation, however, has stagnated mostly owing to spectral-spatial trade-offs. Commonly, low-brightness thermal sources and synchrotron sources are used for Fourier-transform infrared (FTIR) microspectroscopy. Synchrotron sources yield stable, broadband and high-brightness radiation, making them excellent for FTIR microspectroscopy, but the flux of conventional single-beam beamlines is limited by the relatively small horizontal collection angle and the resulting comparatively small source étendue makes them challenging to use with wide-field imaging characterized by a relatively large acceptance or étendue (**Supplementary Note 1**). Here we used multiple synchrotron beams with a wide-field detection scheme. This allowed us to acquire truly diffraction-limited, high-spatial-resolution infrared images of high spectral quality with outstanding speed, considerably extending the potential of infrared microscopy.

For an optical system permitting diffraction-limited imaging, spatial resolution is defined as the capacity to separate two adjacent (point-like) objects. To achieve the highest (diffraction-limited) resolution, an objective with the largest possible numerical aperture (NA) should be used, and the instrument's signal-to-noise ratio (SNR)[4,5] should be optimized. Also, it is indispensable to match the image pixilation to the NA of the objective using the appropriate spatial sampling or pixel size. Too-large pixels inevitably lead to resolution loss, whereas smaller pixels do not improve the resolution further. A detailed analysis[4] (Online Methods) shows that, assuming the largest commercially available NA of ~0.65, diffraction-limited resolution over the entire mid-infrared spectrum can only be achieved with an effective pixel spacing not larger than ~$\lambda/4$ or ~0.6 µm for the shortest wavelength of interest ($\lambda = 2.5$ µm).

One approach to infrared microscopy uses a single element detector and confocal-like apertures to localize light incident on the sample. In this configuration, pixel size is given by the raster-scanning step size[4]. Apertures of dimension *a* only deliver

[1]Department of Physics, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, USA. [2]Synchrotron Radiation Center, University of Wisconsin–Madison, Stoughton, Wisconsin, USA. [3]Department of Bioengineering, Micro and Nanotechnology Laboratory and Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana–Champaign, Urbana, Illinois, USA. [4]Scientific Answers and Solutions, Mount Sinai, New York, USA. [5]Department of Pathology, University of Illinois at Chicago, Chicago, Illinois, USA. Correspondence should be addressed to R.B. (rxb@illinois.edu) or C.J.H. (cjhirsch@uwm.edu).

RECEIVED 18 OCTOBER 2010; ACCEPTED 22 FEBRUARY 2011; PUBLISHED ONLINE 20 MARCH 2011; DOI:10.1038/NMETH.1585

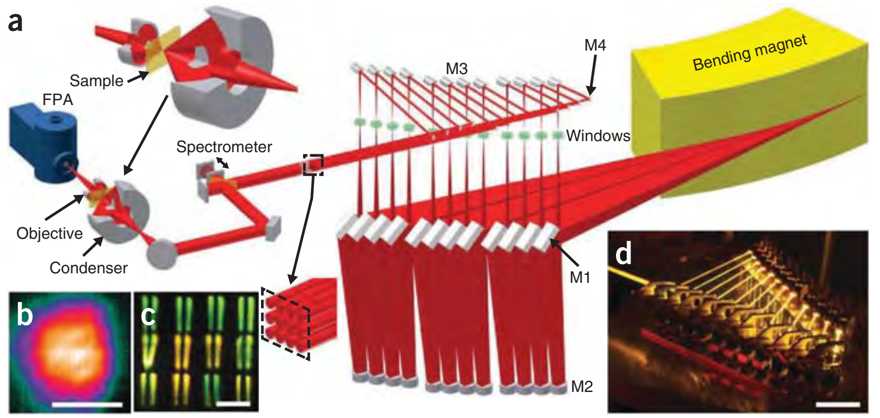


Figure 1 | FTIR imaging with a multibeam synchrotron source. (**a**) Schematic of the experimental setup. M1–M4 are mirror sets. (**b**) A full 128 × 128 pixel FPA image with 12 overlapping beams illuminating an area of ~50 μm × 50 μm. Scale bar, 40 μm. (**c**) A visible-light photograph of the 12 beams projected on a screen in the beam path (dashed box in **a**). Scale bar, ~1.5 cm. We display the beams as one beam from then on in the schematics. Each beam exhibits a shadow cast by a cooling tube upstream, which is not shown in **a**. (**d**) Long-exposure photograph showing the combination of the 12 individual beams into the beam bundle by mirrors M3 and M4. Scale bar, ~20 cm.

diffraction-limited resolution[6] when $\lambda \geq a$. For $\lambda < a$, diffraction-limited resolution[6] is not attained, whereas for longer wavelengths the throughput decays rapidly. This trade-off between resolution and throughput (or SNR) is particularly penalizing for infrared microspectroscopy because of the broad bandwidth. In practice, reasonable SNR limits the smallest aperture for the illumination at the sample plane to ~10 μm × 10 μm for a thermal source[6] and, in a few demonstrations[7], down to ~3 μm × 3 μm for synchrotron sources. The small aperture transmissivity of only a few percent makes point-by-point sampling systems very inefficient because of the dual need for signal averaging to obtain high SNR and rastering a small pixel size to acquire data, leading to exceedingly long acquisition times. These trade-offs make sequential point sampling impractical for micrometer-scale aperture sizes and sub-micrometer-scale raster step sizes (necessary for correct spatial sampling[4]) to achieve diffraction-limited maps. For example, it takes 2–4 h to acquire an area of only 30 μm × 30 μm as a fully diffraction-limited map at a state-of-the-art third-generation synchrotron[7] equipped with a conventional confocal system. Lengthy collection times, in most practical cases, lead experimenters to choose larger aperture and step sizes, thereby compromising the achievable spatial resolution. In contrast, our system can cover this area in under a minute without compromising the spatial sampling required for diffraction-limited resolution.

We based our approach on the more recent strategy of wide-field imaging using multichannel focal plane array (FPA) detectors[8–10], in which no lossy apertures are used. This increases spatial coverage and imaging speed greatly, but the SNR using a thermal source limits pixel sizes to ~5 μm × 5 μm at the sample plane. Achieving a pixel size ~100 times smaller to correctly sample the diffraction-limited illumination is very ineffective, resulting in a

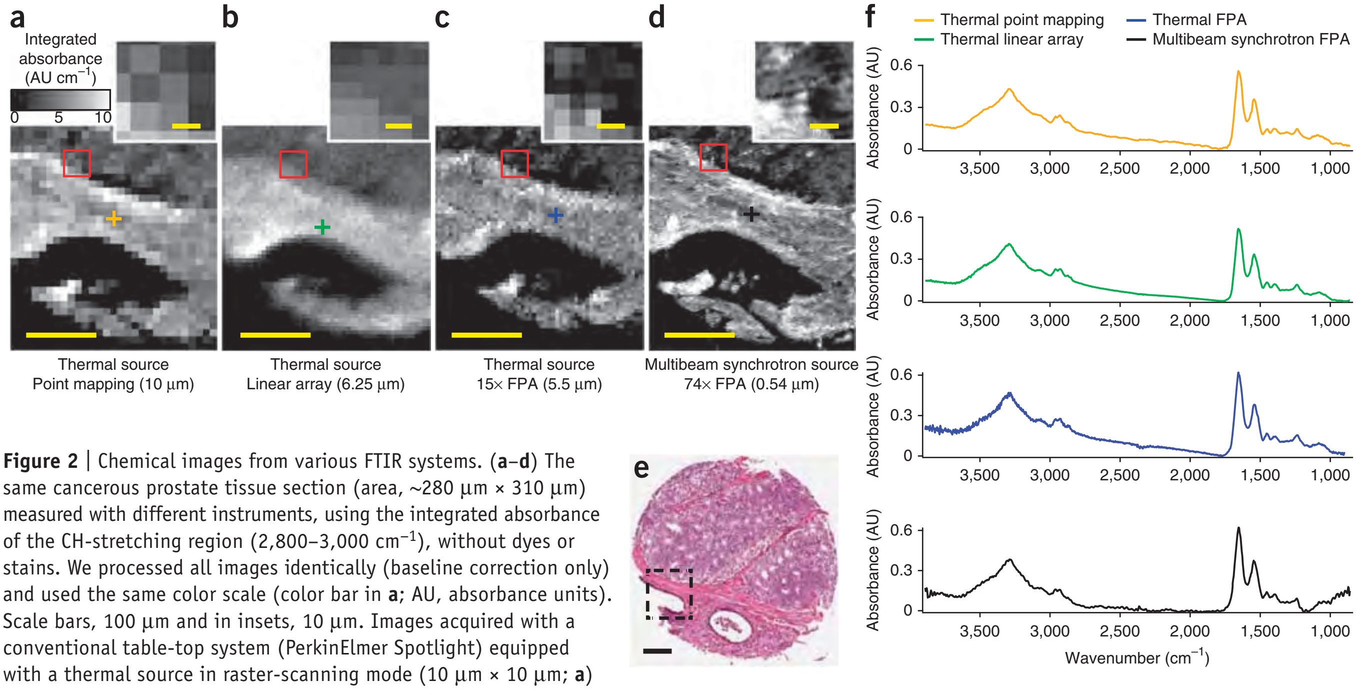


Figure 2 | Chemical images from various FTIR systems. (**a**–**d**) The same cancerous prostate tissue section (area, ~280 μm × 310 μm) measured with different instruments, using the integrated absorbance of the CH-stretching region (2,800–3,000 cm^{-1}), without dyes or stains. We processed all images identically (baseline correction only) and used the same color scale (color bar in **a**; AU, absorbance units). Scale bars, 100 μm and in insets, 10 μm. Images acquired with a conventional table-top system (PerkinElmer Spotlight) equipped with a thermal source in raster-scanning mode (10 μm × 10 μm; **a**) and linear array mode (6.25 μm × 6.25 μm; **b**), with an FTIR imaging system (Varian Stingray) equipped with a 64 pixel × 64 pixel FPA (5.5 × 5.5 μm per pixel at the sample plane; **c**) and with our multibeam synchrotron-based imaging system (pixel size, 0.54 μm × 0.54 μm; **d**). (**e**) Hematoxylin and eosin (H&E)-stained prostate tissue (diameter, 0.75 mm). Scale bar, 100 μm. Dashed box specifies the corresponding area of a serial, unstained section from which we generated images in **a**–**d**. (**f**) Typical unprocessed spectra from a single pixel acquired with each instrument (crosshairs in **a**–**d** indicate corresponding pixel positions in the infrared images).

~100-fold lower SNR (**Supplementary Fig. 1**) and thus in a ~10^4-fold longer scanning time[8]. Hence, to our knowledge there are no reports of a true diffraction-limited FTIR imaging system with a thermal source.

In 2006 independent groups[11–13] pioneered the coupling of a synchrotron beam with an FPA detector, which is not obvious because wide-field illumination seems incompatible with a small, low-emittance synchrotron beam. These groups demonstrated that, with a single synchrotron beam, a local region of the FPA can be illuminated, and that this region yielded increased SNR compared to thermal sources. This inhomogeneous illumination, however, means that either a relatively small FPA (and thus sample area) must be used or that the acquisition time must be increased to compensate the inhomogeneous illumination. This coverage-SNR trade-off has hampered the use of synchrotron-based technology: only one recent publication[14] uses a single synchrotron beam with an FPA.

Here we present an infrared imaging system specifically designed and optimized to overcome these limitations by coupling multiple low-emittance synchrotron beams with a large FPA detector. We extracted a large fan of radiation from a dedicated bending magnet, split it into 12 beams and subsequently rearranged these into a 3 × 4 matrix beam bundle to illuminate a large field of view in the sample plane (**Fig. 1**). We engineered the matrix to achieve homogeneous illumination over areas of up to 52 μm × 52 μm (96 pixels × 96 pixels; **Fig. 1b** and **Supplementary Fig. 2**) with each pixel corresponding to 0.54 μm × 0.54 μm at the sample plane. This pixel size, ~100 times smaller than conventional thermal or synchrotron systems, is smaller than the maximum pixel size allowed for correct spatial sampling (oversampling) so that diffraction-limited images even at the smallest wavelength of interest (2.5 μm) are possible (Online Methods). Although we designed this system explicitly for acquisition in transmission mode, it also yields equivalent quality images in reflection mode (**Supplementary Figs. 3** and **4**).

To test this approach, we compared data from the same prostate tissue using various state-of-the-art infrared imaging systems

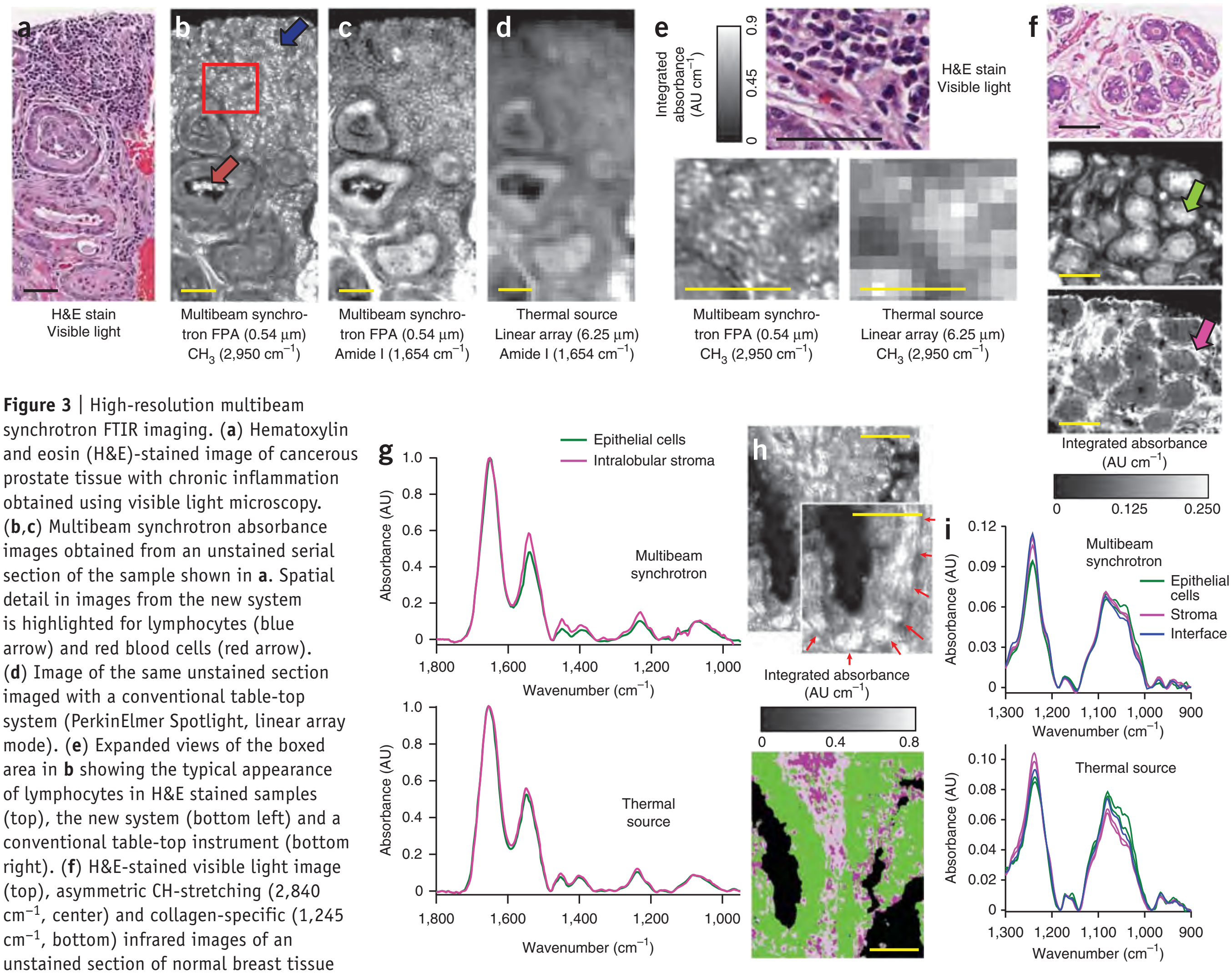


Figure 3 | High-resolution multibeam synchrotron FTIR imaging. (**a**) Hematoxylin and eosin (H&E)-stained image of cancerous prostate tissue with chronic inflammation obtained using visible light microscopy. (**b**,**c**) Multibeam synchrotron absorbance images obtained from an unstained serial section of the sample shown in **a**. Spatial detail in images from the new system is highlighted for lymphocytes (blue arrow) and red blood cells (red arrow). (**d**) Image of the same unstained section imaged with a conventional table-top system (PerkinElmer Spotlight, linear array mode). (**e**) Expanded views of the boxed area in **b** showing the typical appearance of lymphocytes in H&E stained samples (top), the new system (bottom left) and a conventional table-top instrument (bottom right). (**f**) H&E-stained visible light image (top), asymmetric CH-stretching (2,840 cm^{-1}, center) and collagen-specific (1,245 cm^{-1}, bottom) infrared images of an unstained section of normal breast tissue (terminal ductal lobular unit region). Epithelial (green arrow) and intralobular stromal regions (magenta arrow) are highlighted. (**g**) Spectra of epithelial and stromal cells recorded with a multibeam synchrotron versus a thermal source. (**h**) Absorbance image (2,840 cm^{-1}; top) of an unstained cancerous prostate tissue showing two benign prostate glands. Inset, potential presence of basement membrane at the interface between stroma and epithelium is marked (arrows). Image (bottom) showing epithelial (green) and stromal (magenta) cells classified using previous algorithms. (**i**) Average spectra from epithelial, stromal (two each: one closer to the interface, one farther away), and interface pixels identified manually from data obtained using two different instruments. AU, absorbance units. Scale bars, 50 μm.

(**Fig. 2** and **Supplementary Fig. 1**). None of the other instruments provided diffraction-limited resolution at all wavelengths (**Fig. 2a–c**). Raster-scanning the area shown in **Figure 2a–d** (~280 μm × 310 μm) at diffraction-limited resolution using a synchrotron-based dual-aperture microscope would require over 11 d. In contrast, using our technique we recorded the same area (**Fig. 2d**) in ~30 min (16 scans). The spectral quality was essentially identical (**Fig. 2f**) to that of the best commercial systems, despite the ~100-fold pixel area reduction. This pixel size provided the additional spatial detail (**Fig. 2**) necessary for infrared imaging to become competitive with optical microscopy in biomedical applications. In another example, wide-field multibeam synchrotron imaging revealed lymphocytes (diameter, ~2–7 μm) and other tissue features that were clearly visible in hematoxylin and eosin–stained images (the clinical gold standard for diagnosis; **Fig. 3a–c**). The same visualizations were impossible using conventional table-top infrared systems (**Fig. 3d**,**e**). The contrast in these images can be used to color-code images into constituent cell types[15]; hence the capability of our technique opens up the possibility of subcellular classification.

Furthermore, pixel localization also improved spectral purity of data extracted from images. The hematoxylin and eosin contrast was well-reproduced with our technique using simple absorption features, and epithelial and stromal regions were clearly delineated without staining (**Fig. 3f**). The additional detail in synchrotron wide-field images allowed relatively limited cross-contamination of spectra from both intralobular stromal and epithelial regions. Although we expected these characteristic spectra to be different, the limited pixel size of the thermal source systems demonstrated substantial overlap, but the multibeam synchrotron system provided distinct spectra (**Fig. 3g**). Using our technique, we also classified an infrared image of prostate tissue into constituent cell types (**Fig. 3h**). Although it is well-known that the basement membrane lies at the interface of epithelial and stromal cells and is critical in diagnosing lethal cancer, the basement membrane is not discernable in images from thermal systems. We classified infrared tissue images into cell types[15], and identified the interface between the epithelial and stromal cells (**Fig. 3h**). Thermal source spectra from these regions were an average of epithelial and stromal pixels, whereas interface spectra extracted from the synchrotron image were distinct from both contributions (**Fig. 3i**), which, with the higher collagen triplet absorption, was suggestive of the basement membrane. Additional investigations are in progress.

To validate the optical capability of our system, we recorded images of a 1951 US Air Force test target[5] (**Supplementary Figs. 3a**,**b** and **4**). We used line profiles[5] (**Supplementary Fig. 3e–h**) to determine the contrast for each pattern, quantitatively confirming that our system reached and exceeded (**Supplementary Note 2**) the Rayleigh resolution criterion and delivered diffraction-limited images over the entire mid-infrared bandwidth. Furthermore, spatial oversampling at all wavelengths and high SNR, as offered by our system, are a prerequisite[12,13] for developing computational resolution enhancement techniques. We implemented a spatial deconvolution algorithm (**Supplementary Note 3**) based on (wavelength-dependent) measured point-spread functions (**Supplementary Figs. 5** and **6**). The increased contrast and resolution of the deconvolved US Air Force target sample images were apparent in the line profiles (**Supplementary Fig. 3c–h**). Furthermore, measurements of ~1 μm polystyrene beads confirmed that our system reached a spectral limit of detection of 6 ± 1 fmol (mass, 600 ± 100 fg; and volume, 0.6 ± 0.1 fl) in a single 0.54 μm × 0.54 μm pixel (**Supplementary Fig. 7**). We estimated that this limit is about two orders of magnitude finer than that of present instrumentation[16].

The use of multiple synchrotron beams enabled us to achieve a homogeneously high SNR over a large FPA area, which improved sample coverage and acquisition speed compared to conventional thermal or synchrotron-based systems and enabled high diffraction-limited spatial resolution over the entire mid-infrared spectrum. The improvement in acquisition time opens the way to real-time noninvasive and label-free live-cell imaging. We hope that our technique spurs the community to develop appropriate optical designs for table-top instruments and provides a rationale for laser-based systems and other multibeam synchrotron-based imaging beamlines.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank T. Kubala, S. Janowski and M. Fisher for their engineering work, and Z. El-Bayyari for his help during alignment of the beamline. This work was supported by the US National Science Foundation under awards CHE-0832298, CHE-0957849 and DMR-0619759, and by the Research Growth Initiative of the University of Wisconsin–Milwaukee. Part of this work is based on research conducted at the Synchrotron Radiation Center, University of Wisconsin–Madison, which is supported by the National Science Foundation under award DMR-0537588. The project described was also supported by award R01CA138882 from the US National Institutes of Health.

AUTHOR CONTRIBUTIONS

M.J.N., R.R. and C.J.H. designed research; M.J.N., M.J.W. and E.C.M. performed research; M.J.W., A.K.-B., V.M. and R.B. contributed prostate samples; M.J.N., M.J.W., E.C.M., R.B. and C.J.H. analyzed data; and M.J.N., R.B. and C.J.H. wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturemethods/.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Wetzel, D.L. & LeVine, S.M. *Science* **285**, 1224–1225 (1999).
2. Pezacki, J.P. *et al. Nat. Chem. Biol.* **7**, 137–145 (2011).
3. Chalmers, J.M. & Griffiths, P.R. *Handbook of Vibrational Spectroscopy* (Wiley, 2002).
4. Stelzer, E.H.K. *J. Microsc.* **189**, 15–24 (1998).
5. Lasch, P. & Naumann, D. *Biochim. Biophys. Acta* **1758**, 814–829 (2006).
6. Carr, G.L. *Rev. Sci. Instrum.* **72**, 1613–1619 (2001).
7. Dumas, P., Jamin, N., Teillaud, J.L., Miller, L.M. & Beccard, B. *Faraday Discuss.* **126**, 289–302 (2004).
8. Bhargava, R. & Levin, I.W. *Anal. Chem.* **73**, 5157–5167 (2001).
9. Lewis, E.N. *et al. Anal. Chem.* **67**, 3377–3381 (1995).
10. Kidder, L.H., Levin, I.W., Lewis, E.N., Kleiman, V.D. & Heilweil, E.J. *Opt. Lett.* **22**, 742–744 (1997).
11. Moss, D., Gasharova, B. & Mathis, Y. *Infrared Phys. Technol.* **49**, 53–56 (2006).
12. Carr, G.L., Chubar, O. & Dumas, P. in *Spectrochemical Analysis Using Infrared Multichannel Detectors* 1st edn. (eds., Bhargava, R. & Levin, I.W.) 56–84 (Wiley-Blackwell, Oxford, 2005).
13. Miller, L.M. & Dumas, P. *Biochim. Biophys. Acta* **1758**, 846–857 (2006).
14. Petibois, C., Cestelli-Guidi, M., Piccinini, M., Moenner, M. & Marcelli, A. *Anal. Bioanal. Chem.* **397**, 2123–2129 (2010).
15. Fernandez, D.C., Bhargava, R., Hewitt, S.M. & Levin, I.W. *Nat. Biotechnol.* **23**, 469–474 (2005).
16. Bhargava, R., Schwartz Perlman, R., Fernandez, D.C., Levin, I.W. & Bartick, E.G. *Anal. Bioanal. Chem.* **394**, 2069–2075 (2009).

ONLINE METHODS

Requirements for diffraction-limited resolution. Mid-infrared spectroscopy and microscopy has very broad applications in many scientific fields, ranging from fundamental and applied research to engineering and biology[15–29]. Infrared microspectroscopy in particular can contribute to the biomedical sciences because of its noninvasive spatially resolved chemical specificity. Here we describe the requirements to obtain diffraction-limited spatial resolution with a mid-infrared microscope.

Spatial resolution can be quantified, for example, by the Rayleigh[5] criterion as $d = 0.61\ \lambda\ /\ \mathrm{NA}$, in which d is the minimum distance between two adjacent (point-like) objects that are just resolved (the factor 0.61 is strictly valid only for lenses without obscuration and smaller for Schwarzschild optics; see **Supplementary Note 2**). But achievable spatial resolution is not only dependent on the wavelength and the NA of the objective via the Rayleigh criterion but also on the pixel size, that is, the objective's magnification and the SNR of the imaging system[4]. To observe diffraction-limited performance, a spatial sampling of at least ~8 pixels[4] per Airy pattern is required to achieve sufficient contrast. Smaller pixel sizes (oversampling) do not improve the resolution, which is then limited by diffraction, whereas larger pixels unavoidably deteriorate contrast and thus resolution (undersampling). For the smallest wavelength (2.5 μm) using an NA of 0.65, we need a pixel size not larger than 1.22 × 2.5 μm / 0.65 / 8 = 0.59 μm. Even the less restrictive Nyquist theorem yields a maximum pixel size of $1\ /\ (2.3\ f_{\mathrm{cutoff}}) = 0.84$ μm (usually 2.3 is used instead of the theoretical 2 suggested by Nyquist to account for factors such as noise in real optical systems[30]), where $f_{\mathrm{cutoff}} = 2\ \mathrm{NA}\ /\ \lambda$ is the spatial cutoff frequency, equivalent to the Sparrow frequency[5,31]. In summary, this means that the NA of an objective alone is not enough to provide the resolution promised by the Rayleigh criterion, but its magnification also has to match. In the case of an objective with an NA of 0.65 (approximately the largest commercially available NA, giving the best possible spatial resolution), it needs at least a magnification of 40 μm / 0.59 μm = 68 (assuming a typical FPA pixel size of 40 μm × 40 μm). We used a 74× objective (NA = 0.65) in our setup, leading to a pixel size of 0.54 μm × 0.54 μm (slight oversampling). In addition this high spatial sampling offers the advantage that subdiffraction objects can be localized (but of course, not resolved) with an accuracy better than the diffraction limit[32].

Instrument design. Synchrotron storage rings are excellent light sources for aperture-based infrared microspectroscopy[33] as the small horizontal and vertical emittance (source étendue) of conventional single-beam beamlines and the relatively small acceptance (detector system étendue) of the microscopy system can be closely matched (**Supplementary Table 1**). Increasing the photon flux by extracting a larger horizontal angle from a bending magnet, however, is not beneficial because the additional photons cannot be coupled efficiently to the small acceptance of such microscopy systems. For wide-field microscopes without throughput-restricting apertures, in contrast, single beams from conventional beamlines have limited flux owing to their relatively small emittance, making it challenging to match the relatively large acceptance of a multichannel FPA imaging instrument. The instrument described here substantially increased the horizontal collection angle to match the large acceptance of a wide-field imaging system to fully exploit the source brightness. It is located at the Synchrotron Radiation Center in Stoughton, Wisconsin, USA, which already houses a conventional aperture-based infrared microscope. This synchrotron facility encourages scientists to apply for peer-reviewed access to beamtime and/or initiate a collaboration with the authors of this work. Applications are accepted for review every six months and rapid requests for initial experiments are handled more frequently (http://www.src.wisc.edu/users/new_users.html).

We extracted 320 mrad × 27 mrad of infrared radiation from a dedicated bending magnet and split this fan of radiation into twelve beams with a set of twelve toroidal mirrors (M1; **Fig. 1**), which refocused each beam (magnification of 1). Each beam exited an ultrahigh vacuum chamber via one of twelve flat mirrors (M2; **Fig. 1**) through one of twelve ZnSe windows (**Fig. 1**) into a nitrogen-purged area. Next, twelve parabolic mirrors (M3; **Fig. 1**) collimated the beams, followed by twelve stacked small flat mirrors (M4; **Fig. 1**) that rearranged the beams into a 3 × 4 matrix. We used a subsequent piezo-driven optical feedback system (feedback system is not shown) to stabilize the beam bundle, reduce vibration effects and increase the SNR. Next, we sent the bundle through a Vertex 70 (Bruker) spectrometer (**Fig. 1**), which was coupled to a Hyperion 3000 (Bruker) infrared and visible light microscope. There, the slightly defocused beam bundle illuminated the sample area through a 15× or 20× Schwarzschild condenser (**Fig. 1**) to spread out each beam so that the beams overlap spatially to provide quasi-homogeneous illumination at the sample. Finally, a 74× objective (Ealing) imaged the sample onto a 128 pixel × 128 pixel FPA (Santa Barbara Focalplane), so that each pixel had an effective geometrical area at the sample plane of 0.54 μm × 0.54 μm (**Fig. 1**). Additional design details of the imaging system have been reported elsewhere[34]. In contrast to other implementations of thermal or synchrotron sources, our multibeam system allowed us to simultaneously uniformly illuminate an order of magnitude more pixels (96 pixels × 96 pixels; **Fig. 1b**) and used an objective with a substantially higher NA of 0.65 with a correctly matched[4] pixel size (0.54 μm × 0.54 μm) to maintain full high diffraction-limited resolution over the mid-infrared spectrum at a high SNR. We used a condenser with an NA of ~0.6 to match the NA of the objective. Owing to its higher NA, this objective delivered 38% and 23% higher spatial resolution (according to the Rayleigh criterion) compared to previous studies (for example, the 15× objective with NA = 0.4 and pixel size = 2.7 μm × 2.7 μm or 36× objective with NA = 0.5 and pixel size = 1.1 μm × 1.1 μm)[11,14]. Furthermore, owing to the multibeam design, a high synchrotron storage ring current was not mandatory to obtain high SNR. The ~270 mA current of our storage ring was sufficient to achieve similar SNR (**Fig. 2d**,**f**) leading to shorter acquisition times compared to those reported in previous publications[14]. The present design can cover more than double the sample area in equivalent or shorter times with better spatial resolution as compared to single synchrotron beam systems.

Synchrotron sources may have coherent properties, for example, synchrotrons with pulse lengths shorter than tens of femtoseconds in the far infrared. The present source, however, had nanosecond pulses, and we designed the path lengths for the twelve beams to never temporally overlap on the sample or detector plane. Hence, temporal coherence did not have an impact on the imaging quality of the images produced by the microscope. Experimentally we observed no spectral evidence of spatial or temporal coherence effects, nor any impact on image quality or resolution, as can be seen, for example, by the correspondence between the thermal and synchrotron spectral data.

Experimental details, **data processing, and samples.** We conducted conventional thermal source-based imaging on two commercial systems: Stingray (Varian; **Fig. 2c**) using an FPA detector and Spotlight 400 (PerkinElmer; **Figs. 2a**,**b** and **3d**,**e** and **Supplementary Fig. 1a**) equipped with a single element and a 16-pixel linear array detector. We acquired the synchrotron point-by-point scanning image (**Supplementary Fig. 1b**) on a Continuμm (Thermo Nicolet) dual-aperture microscope connected to beamline 031, and we collected the remaining images with the multibeam synchrotron system connected to a Hyperion 3000 (Bruker) microscope at beamline 021, both at the Synchrotron Radiation Center. The Varian, PerkinElmer and Thermo Nicolet measurements used a Happ-Genzel, the Bruker measurements a Norton-Beer (medium) apodization. We baseline-corrected the images in **Figures 2** and **3** (including spectra), **Supplementary Figures 1** and **7**; all other infrared images as well as spectra show raw data. We did not use post-acquisition smoothing or filtering. The infrared data were analyzed and images were created with software packages IRidys (in-house development) and ENVI (ITT VIS).

The prostate cancer sample (Gleason grade 6) with epithelial cells (**Fig. 2** and **Supplementary Fig. 1**) was a viable tumor without necrosis, in a cribriforming pattern and had some strands of stroma crossing through it. A second prostate cancer sample, which was also Gleason grade 6 for comparison (**Fig. 3a**–**e**), had chronic inflammation (mostly mononuclear cell infiltration of macrophages and lymphocytes) and contained two glands, a small vessel with a muscular wall and capillaries (with blood). The tissue shown in **Figure 3f** was a normal human breast tissue core including the terminal ductal lobular unit (TDLU) region and the tissue shown in **Figure 3h** contained two benign prostate glands from a cancerous prostate tissue core (Gleason grade 6). Tissues used here were from anonymized samples from individuals and involved secondary analysis as approved by the University of Illinois at Urbana-Champaign Institutional Review Board, protocol 06684. We fixed all biomedical samples in 4% paraformaldehyde, embedded them in paraffin, sectioned them at a thickness of 4 μm, mounted them on a BaF_2 infrared transparent window and deparaffinized them with hexane for 48 h before measurement. In transmission mode sample thickness can affect the obtainable spatial resolution. Using a simple geometric model we estimated that the sample thickness should not be above ~3–4 μm to achieve full diffraction-limited resolution.

We purchased the apertures (**Supplementary Figs. 5** and **6**) from National Aperture, Inc., the high-resolution US Air Force (USAF) test target (**Supplementary Fig. 3**) from Edmund Optics Inc. and the polystyrene beads (**Supplementary Fig. 7**) from Polysciences, Inc. We diluted the polystyrene bead suspension with water, dispensed it on an ultrathin formvar film substrate and then air-dried it.

We recorded images of polystyrene beads with a diameter of ~1 and 2 μm (acquisition time, ~5 min) to examine spectral limits of detection per pixel. We detected the 6 ± 1 fmol or 3.4×10^9 ($\pm 0.7 \times 10^9$; s.d.) CH_2 groups contained in a 1 μm polystyrene bead (mass, 600 ± 100 fg; volume, 0.6 ± 0.1 fl) in a single 0.54 μm × 0.54 μm pixel using the International Union of Pure and Applied Chemistry (IUPAC) detection limit criterion (**Supplementary Fig. 7**). We estimated this to be ~100-fold better than with current instrumentation[16] and this compared favorably with the lowest detection limit reported[35] using destructive methods.

17. Li, Z.Q. *et al. Nat. Phys.* **4**, 532–535 (2008).
18. Bunaciu, A.A., Aboul-Enein, H.Y. & Fleschin, S. *Appl. Spectrosc. Rev.* **45**, 206–219 (2010).
19. Matveev, S. & Stachel, T. *Geochim. Cosmochim. Acta* **71**, 5528–5543 (2007).
20. Prati, S., Joseph, E., Sciutto, G. & Mazzeo, R. *Acc. Chem. Res.* **43**, 792–801 (2010).
21. Politi, Y., Arad, T., Klein, E., Weiner, S. & Addadi, L. *Science* **306**, 1161–1164 (2004).
22. Martin, F.L. *et al. Nat. Protoc.* **5**, 1748–1760 (2010).
23. Hazen, T.C. *et al. Science* **330**, 204–208 (2010).
24. Walsh, M.J. *et al. Stem Cells* **26**, 108–118 (2008).
25. Walsh, M.J. *et al. Stem Cell Res.* **3**, 15–27 (2009).
26. Holman, H.N., Bechtel, H.A., Hao, Z. & Martin, M.C. *Anal. Chem.* **82**, 8757–8765 (2010).
27. Kuzyk, A. *et al. J. Biol. Chem.* **285**, 31202–31207 (2010).
28. Webster, G.T. *et al. Anal. Chem.* **81**, 2516–2524 (2009).
29. Bhargava, R. *Anal. Bioanal. Chem.* **389**, 1155–1169 (2007).
30. Centonze, V. & Pawley, J.B. in *Handbook of Biological Confocal Microscopy* 3rd edn. (ed. Pawley, J.B.) 627–649 (Springer, New York, 2006).
31. Murphy, D.B. *Fundamentals of Light Microscopy and Electronic Imaging* 1st edn. 233–258 (Wiley-Liss, New York, 2001).
32. Bobroff, N. *Rev. Sci. Instrum.* **57**, 1152–1157 (1986).
33. Reffner, J.A., Martoglio, P.A. & Williams, G.P. *Rev. Sci. Instrum.* **66**, 1298–1302 (1995).
34. Nasse, M.J., Reininger, R., Kubala, T., Janowski, S. & Hirschmugl, C. *Nucl. Instrum. Methods Phys. Res. A* **582**, 107–110 (2007).
35. Park, K., Lee, J., Bhargava, R. & King, W.P. *Anal. Chem.* **80**, 3221–3228 (2008).

doi:10.1038/nmeth.1585

On the Homogenization of Data from Two Laboratories Using Genetic Programming

Jose G. Moreno-Torres[1], Xavier Llorà[2], David E. Goldberg[3],
and Rohit Bhargava[4]

[1] Department of Computer Science and Artificial Intelligence,
Universidad de Granada, 18071 Granada, Spain
jose.garcia.mt@decsai.ugr.es
[2] National Center for Supercomputing Applications (NCSA)
University of Illinois at Urbana-Champaign
1205 W. Clark Street, Urbana, Illinois, USA
xllora@illinois.edu
[3] Illinois Genetic Algorithms Laboratory (IlliGAL)
University of Illinois at Urbana-Champaign
104 S. Mathews Ave, Urbana, Illinois, USA
deg@illinois.edu
[4] Department of Bioengineering
University of Illinois at Urbana-Champaign
405 N. Mathews Ave, Urbana, Illinois, USA
rbx@uiuc.edu

Abstract. In experimental sciences, diversity tends to difficult predictive models' proper generalization across data provided by different laboratories. Thus, training on a data set produced by one lab and testing on data provided by another lab usually results in low classification accuracy. Despite the fact that the same protocols were followed, variability on measurements can introduce unforeseen variations that affect the quality of the model. This paper proposes a Genetic Programming based approach, where a transformation of the data from the second lab is evolved driven by classifier performance. A real-world problem, prostate cancer diagnosis, is presented as an example where the proposed approach was capable of repairing the fracture between the data of two different laboratories.

1 Introduction

The assumption that a properly trained classifier will be able to predict the behavior of unseen data from the same problem is at the core of any automatic classification process. However, this hypothesis tends to prove unreliable when dealing with biological data (or other experimental sciences), especially when such data is provided by more than one laboratory, even if they are following the same protocols to obtain it.

This paper presents an example of such a case, a prostate cancer diagnosis problem where a classifier built using the data of the first laboratory performs

J. Bacardit et al. (Eds.): IWLCS 2008/2009, LNAI 6471, pp. 185–197, 2010.

very accurately on the test data from that same laboratory, but comparatively poorly on the data from the second one. It is assumed that this behavior is due to a fracture between the data of the two laboratories, and a Genetic Programming (GP) method is developed to homogenize the data in subsequent subsets. We consider this method a form of feature extraction because the new dataset is constructed with new features which are functional mappings of the old ones.

The method presented in this paper attempts to optimize a transformation over the data from the second laboratory, in terms of classifier performance. That is, the data from the second lab is transformed into a new dataset where the classifier, trained on the data from the first lab, performs as accurately as possible. If the performance achieved by the classifier in this new, transformed, dataset, is equivalent to the one obtained in the data from the first lab, we understand the data has been homogenized.

More formally, the classifier f is trained on data from one laboratory (dataset A), such that $y = f(xA)$ is the class prediction for one instance xA of dataset A. For the data from the other lab (dataset B), it is assumed that there exists a transformation T such that $f(T(xB))$ is a good classifier for instances xB of dataset B. The 'goodness' of the classifier is measured by the loss function $l(f(T(xB)), y)$, where y is the class associated with xB, and $l(.,.)$ is a measure of distance between $f(T(xB))$ and y. The aim is to find a transformation T such that the average loss over all instances in B is minimized.

The remainder of this paper is organized as follows: In Section 2, some preliminaries about the techniques used and some approaches to similar problems in the literature are presented. Section 3 has a description of the proposed algorithm. Section 4 details the real-world biological dataset that motivates this paper. Section 5 includes the experimental setup, along with the results obtained, and an analysis. Finally, some concluding remarks are made in Section 6.

2 Preliminaries

This section is divided in the following way: In Section 2.1 we introduce the notation that has been used in this paper. Then we include a brief summary of what has been done in feature extraction in Section 2.2, and a short review of the different approaches we found in the specialized literature on the use of GP for feature extraction in Section 2.3.

2.1 Notation

When describing the problem, datasets A, B and S correspond to:

- A: The original dataset, provided by the first lab, that was used to build the classifier.
- B: The problem dataset, from the second lab. The classifier is not accurate on this dataset, and that is what the proposed algorithm attempts to solve.
- S: The solution dataset, result of applying the evolved transformation to the samples in dataset B. The goal is to have the classifier performance be as high as possible on this dataset.

2.2 Feature Extraction

Feature extraction is one form of pre-processing, which creates new features as functional mappings of the old ones. An early proposer of such a term was probably Wyse in 1980 [1], in a paper about intrinsic dimensionality estimation. There are multiple techniques that have been applied to feature extraction throughout the years, ranging from principal component analysis (PCA) to support vector machines (SVMs) to GAs (see [2,3,4], respectively, for some examples).

Among the foundations papers in the literature, Liu's book in 1998 [5] is one of the earlier compilations of the field. A workshop held in 2003 [6], led Guyon & Elisseeff to publish a book with an important treatment of the foundations of feature extraction[7].

2.3 Genetic Programming-Based Feature Extraction

Genetic Programming (GP) has been used extensively to optimize feature extraction and selection tasks. One of the first contributions in this line was the work published by Tackett in 1993 [8], who applied GP to feature discovery and image discrimination tasks.

We can consider two main branches in the philosophy of GP-based feature extraction:

1 On one hand, we have the proposals that focus only on the feature extraction procedure, of which there are multiple examples: Sherrah et al. [9] presented in 1997 the evolutionary pre-processor (EPrep), which searches for an optimal feature extractor by minimizing the misclassification error over three randomly selected classifiers. Kotani et al.'s work from 1999 [10] determined the optimal polynomial combinations of raw features to pass to a k-nearest neighbor classifier. In 2001, Bot [11] evolved transformed features, one-at-a-time, again for a k-NN classifier, utilizing each new feature only if it improved the overall classification performance. Zhang & Rockett, in 2006, [12] used multiobjective GP to learn optimal feature extraction in order to fold the high-dimensional pattern vector to a one-dimensional decision space where the classification would be trivial. Lastly, also in 2006, Guo & Nandi [13] optimized a modified Fisher discriminant using GP, and then Zhang & Rockett [14] extended their work by using a multiobjective approach to prevent tree bloat.

2 On the other hand, some authors have chosen to evolve a full classifier with an embedded feature extraction step. As an example, Harris [15] proposed in 1997 a co-evolutionary strategy involving the simultaneous evolution of the feature extraction procedure along with a classifier. More recently, Smith & Bull [16] developed a hybrid feature construction and selection method using GP together with a GA.

2.4 Finding and Repairing Fractures between Data

Among the proposals to quantify the fracture in the data, we would like to mention the one by Wang et al. [17], where the authors present the idea of correspondence tracing. They propose an algorithm for the discovering of changes of classification characteristics, which is based on the comparison between two rule-based classifiers, one built from each dataset. Yang et al. [18] presented in 2008 the idea of conceptual equivalence as a method for contrast mining, which consists of the discovery of discrepancies between datasets. Lately, it is important to mention the work by Cieslak and Chawla [19], which presents a statistical framework to analyze changes in data distribution resulting in fractures between the data.

The fundamental difference between the mentioned works and this one is we focus on repairing the fracture by modifying the data, using a general method that works with any kind of data fracture, while they propose methods to quantify said fracture that work provided some conditions.

3 A Proposal for GP-Based Feature Extraction to Homogenize Data from Two Laboratories

The problem we are attempting to solve is the design of a method that can create a transformation from a dataset (dataset B) where a classification model built using the data from a different dataset (dataset A) is not accurate; into a new dataset (dataset S) where the classifier is more accurate. Said classifier is kept unchanged throughout the process.

We decided to use GP to solve the problem for a number of reasons:

1 It is well suited to evolve arbitrary expressions because its chromosomes are trees. This is useful in our case because we want to have the maximum possible flexibility in terms of the functional expressions of this transformations.
2 GP provides highly-interpretable solutions. This is an advantage because our goal is not only to have a new dataset where the classifier works, but also to analyze what was the problem in the first dataset.

Once GP was chosen, we needed to decide what terminals and operators to use, how to calculate the fitness of an individual and which evolutionary parameters (population size, number of generations, selection and mutation rates, etc) are appropriate for the problem at hand.

3.1 Solutions Representation: Context-Free Grammar

The representation of the solutions was achieved by extending GP to evolve more than one tree per solution. Each individual is composed by n trees, where n is the number of attributes present in the dataset. We are trying to develop a new dataset with the same number of attributes as the old one, since this new dataset needs to be fed to the existing model. In the tree structure, the leaves are either constants (we use the Ephemeral Random Constant approach [20]) or attributes from the original dataset. The intermediate nodes are functions from the function set, which is specific to each problem.

The attributes on the transformed dataset are represented by algebraic expressions. These expressions are generated according to the rules of a context-free grammar which allows the absence of some of the functions or terminals. The grammar corresponding to the example problem would look like this:

$$\begin{aligned}
&Start \rightarrow Tree\,Tree\\
&Tree \rightarrow Node\\
&Node \rightarrow Node\;Operator\;Node\\
&Node \rightarrow Terminal\\
&Operator \rightarrow + \mid - \mid * \mid \div\\
&Terminal \rightarrow x_0 \mid x_1 \mid E\\
&E \rightarrow realNumber(represented\;by\;e)
\end{aligned}$$

3.2 Fitness Evaluation

The fitness evaluation procedure is probably the most treated aspect of design in the literature when dealing with GP-based feature extraction. As has been stated before, the idea is to have the provided classifier's performance drive the evolution. To achieve that, our method calculates fitness as the classifier's accuracy over the dataset obtained by applying the transformations encoded in the individual (training-set accuracy).

3.3 Genetic Operators

This section details the choices made for selection, crossover and mutation operators. Since the objective of this work is not to squeeze the maximum possible performance from GP, but rather to show that it is an appropriate technique for the problem and that it can indeed solve it, we did not pay special attention to these choices, and picked the most common ones in the specialized literature.

- Tournament selection without replacement. To perform this selection, s individuals are first randomly picked from the population (where s is the tournament size), while avoiding using any member of the population more than once. The selected individual is then chosen as the one with the best fitness among those picked in the first stage.
- One-point crossover: A subtree from one of the parents is substituted by one from the other parent. This procedure is carried over in the following way:

  1 Randomly select a non-root non-leave node on each of the two parents.
  2 The first child is the result of swapping the subtree below the selected node in the father for that of the mother.
  3 The second child is the result of swapping the subtree below the selected node in the mother for that of the father.

- Swap mutation: This is a conservative mutation operator, that helps diversify the search within a close neighborhood of a given solution. It consists of exchanging the primitive associated to a node by one that has the same number of arguments.
- Replacement mutation: This is a more aggressive mutation operator that leads to diversification in a larger neighborhood. The procedure to perform this mutation is the following:
  1 Randomly select a non-root non-leave node on the tree to mutate.
  2 Create a random tree of depth no more than a fixed maximum depth. In this work, the maximum depth allowed was 5.
  3 Swap the subtree below the selected node for the randomly generated one.

3.4 Function Set

Which functions to include in the function set are usually dependent on the problem. Since one of our goals is to have an algorithm as universal and robust as possible, where the user does not need to fine-tune any parameters to achieve good performance; we decided not to study the effect of different function set choices. We chose the default functions most authors use in the literature: $\{+, -, *, \div, exp, cos\}$.

3.5 Parameters

Table 1 summarizes the parameters used for the experiments.

Table 1. Evolutionary parameters for a n_v-dimensional problem

Parameter	Value
Number of trees	n_v
Population size	$400 * n_v$
Duration of the run	100 generations
Selection operator	Tournament without replacement
Tournament size	$log_2(n_v) + 1$
Crossover operator	One-point crossover
Crossover probability	0.9
Mutation operator	Replacement & Swap mutations
Replacement mutation probability	0.001
Swap mutation probability	0.01
Maximum depth of the swapped in subtree	5
Function set	$\{+, -, *, \div, cos, exp\}$
Terminal set	{x_0,x_1,...,$x_{n_v} - 1$, e}

3.6 Execution Flow

Algorithm 1 contains a summary of the execution flow of the GP procedure, which follows a classical evolutionary scheme. It stops after a user-defined number of generations,

Algorithm 1. Execution flow of the GP method

```
1. Randomly create the initial population by applying the
   context-free grammar in Section 3.1.
2. Repeat Ng times (where Ng is the number of generations)
  2.1 Evaluate the current population, using the procedure
       seen in Section 3.2.
  2.2 Apply selection and crossover to create a new
       population that will replace the old one.
  2.3 Apply the mutation operators to the new population.
3. Return the best individual ever seen.
```

4 Case Study: Prostate Cancer Diagnosis

Prostate cancer is the most common non-skin malignancy in the western world. The American Cancer Society estimated 192,280 new cases and 27,360 deaths related to prostate cancer in 2009 [21]. Recognizing the public health implications of this disease, men are actively screened through digital rectal examinations and/or serum prostate specific antigen (PSA) level testing. If these screening tests are suspicious, prostate tissue is extracted, or biopsied, from the patient and examined for structural alterations. Due to imperfect screening technologies and repeated examinations, it is estimated that more than one million people undergo biopsies in the US alone.

4.1 Diagnostic Procedure

Biopsy, followed by manual examination under a microscope is the primary means to definitively diagnose prostate cancer as well as most internal cancers in the human body. Pathologists are trained to recognize patterns of disease in the architecture of tissue, local structural morphology and alterations in cell size and shape. Specific patterns of specific cell types distinguish cancerous and non-cancerous tissues. Hence, the primary task of the pathologist examining tissue for cancer is to locate foci of the cell of interest and examine them for alterations indicative of disease. A detailed explanation of the procedure is beyond the scope of this paper and can be found elsewhere [22,23,24,25].

Operator fatigue is well-documented and guidelines limit the workload and rate of examination of samples by a single operator (examination speed and throughput). Importantly, inter- and intra-pathologist variation complicates decision making. For this reason, it would be extremely interesting to have an accurate automatic classifier to help reduce the load on the pathologists. This was partially achieved in [24], but some issues remain open.

4.2 The Generalization Problem

Llorà et al. [24] successfully applied a genetics-based approach to the development of a classifier that obtained human-competitive results based on FTIR data. However, the classifier built from the data obtained from one laboratory proved remarkably inaccurate when applied to classify data from a different hospital. Since all the experimental procedure was identical; using the same machine, measuring and post-processing; and having the exact same lab protocols, both for tissue extraction and staining; there was no factor that could explain this discrepancy.

What we attempt to do with this work is develop an algorithm that can evolve a transformation over the data from the second laboratory, creating a new dataset where the classifier built from the first lab is as accurate as possible.

4.3 Pre-processing of the Data

The biological data obtained from the laboratories has an enormous size (in the range of 14GB of storage per sample); and parallel computing was needed to achieve better-than-human results. For this reason, feature selection was performed on the dataset obtained by FTIR. It was done by applying an evaluation of pairwise error and incremental increase in classification accuracy for every class, resulting in a subset of 93 attributes. This reduced dataset provided enough information for classifier performance to be rather satisfactory: a simple C4.5 classifier achieved $\sim 95\%$ accuracy on the data from the first lab, but only $\sim 80\%$ on the second one. The dataset consists of 789 samples from one laboratory and 665 from the other one. These samples represent 0.01% of the total data available for each data set, which were selected applying stratified sampling without replacement. A detailed description of the data pre-processing procedure can be found in [22].

The experiments reported in this paper were performed utilizing the reduced dataset, since the associated computational costs make it unfeasible to work with the complete one. The reduced dataset is made of 93 real attributes, and there are two classes (positive and negative diagnosis). The dataset consists of 789 samples from one laboratory and 665 from the other one, with a $60\% - 40\%$ class distribution.

5 Experimental Study

This section is organized in the following way: To begin with, a general description of the experimental procedure is presented in Section 5.1, and the parameters used for the experiment. The results obtained are presented in Section 5.2, a statistical analysis is shown in Section 5.3, and lastly some sample transformations are shown in Section 5.4.

5.1 Experimental Framework

The experimental methodology can be summarized as follows:

1 Consider each of the provided datasets (one from each lab) to be datasets A and B respectively.

2 From dataset A, build a classifier. We chose C4.5 [26], but any other classifier would work exactly the same; due to the fact that the proposed method uses the learned classifier as a black box.
3 Apply our method to dataset B in order to evolve a transformation that will create a solution dataset S. Use 5-fold cross validation over dataset S, so that training and test set accuracy results can be obtained.
4 Check the performance of the step 2 classifier on dataset S. Ideally, it should be close to the one on dataset A, meaning the proposed method has successfully discovered the hidden transformation and inverted it.

5.2 Performance Results

This section presents the results for the Prostate Cancer problem, in terms of classifier accuracy. The results obtained can be seen in table 2.

Table 2. Classifier performance results

Classifier performance in dataset ...				
A-training	A-test	B	S-training	S-test
0.95435	0.92015	0.83570	0.95191	0.92866

The performance results are promising. First and foremost, the proposed method was able to find a transformation over the data from the second laboratory that made the classifier work just as well as it did on the data from the first lab, effectively finding the fracture in the data (that is, the difference in data distribution between the data sets provided by the two labs) that prevented the classifier from working accurately.

5.3 Statistical Analysis

To complete the experimental study, we performed a statistical comparison between the classifier performance over datasets A, B and S.

In [27,28,29,30] a set of simple, safe and robust non-parametric tests for statistical comparisons of classifiers are recommended. One of them is the Wilcoxon Signed-Ranks Test [31,32], which is the test that we have selected to do the comparison.

In order to perform the Wilcoxon test, we used the results from each partition in the 5-fold cross validation procedure. We ran the experiment four times, resulting in $4 * 5 = 20$ performance samples to carry out the statistical test. R^+ corresponds to the first algorithm in the comparison winning, R^- to the second one.

We can conclude our method has proved to be capable of fully homogenizing the data from both laboratories regarding classifier performance, both in terms of training and test set.

Table 3. Wilcoxon signed-ranks test results

Comparison	R^+	R^-	p-value	null hypothesis of equality
A-test vs B	210	0	$1.91E-007$	*rejected* (A-test outperforms B)
B vs S-test	0	210	$1.91E-007$	*rejected* (S-test outperforms B)
A-training vs S-training	126	84	$--$	*accepted*
A-test vs S-test	84	126	$--$	*accepted*

5.4 Obtained Transformations

Figure 1 contains a sample of some of the evolved expressions for the best individual found by our method. Since the dataset has 93 attributes, the individual was composed of 93 trees, but for space concerns only the attributes relevant to the C4.5 classifier were included here.

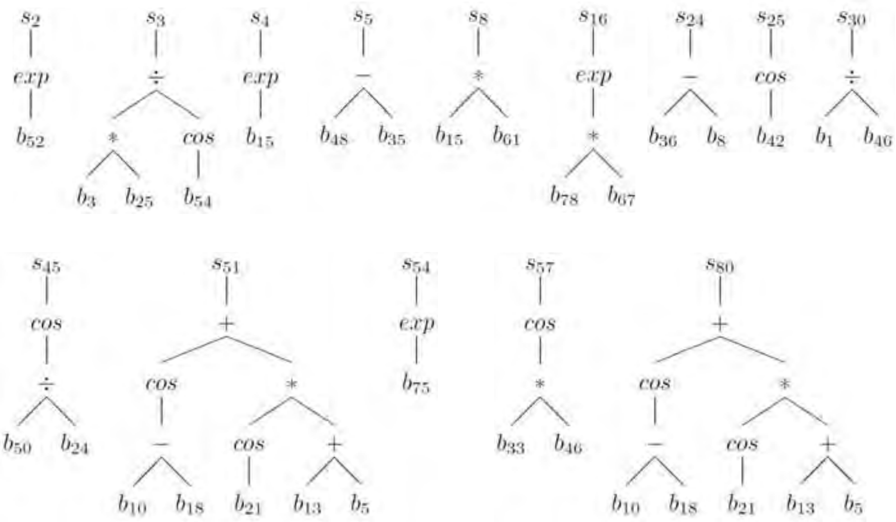


Fig. 1. Tree representation of the expressions contained in a solution to the Prostate Cancer problem

6 Concluding Remarks

We have presented a new algorithm that approaches a common problem in real life for which not many solutions have been proposed in evolutionary computing. The problem in question is the repairing of fractures between data by adjusting the data itself, not the classifiers built from it.

We have developed a solution to the problem by means of a GP-based algorithm that performs feature extraction on the problem dataset driven by the accuracy of the previously built classifier.

We have applied our method to a real-world problem where data from two different laboratories regarding prostate cancer diagnosis was provided, and where the classifier learned from one did not perform well enough on the other. Our algorithm was capable of learning a transformation over the second dataset that made the classifier fit just as well as it did on the first one. The validation results with 5-fold cross validation also support the idea that the algorithm is obtaining good results; and has a strong generalization power.

We have applied a statistical analysis methodology that supports the claim that the classifier performance obtained on the solution dataset significantly outperforms the one obtained on the problem dataset.

Lastly, we have shown the learned transformations. Unfortunately, we have not been able to extract any useful information from them yet.

Acknowledgments

Jose García Moreno-Torres was supported by a scholarship from 'Obra Social la Caixa' and is currently supported by a FPU grant from the Ministerio de Educación y Ciencia of the Spanish Government and the KEEL project. Rohit Bhargava would like to acknowledge collaborators over the years, especially Dr. Stephen M. Hewitt and Dr. Ira W. Levin of the National Institutes of Health, for numerous useful discussions and guidance. Funding for this work was provided in part by University of Illinois Research Board and by the Department of Defense Prostate Cancer Research Program. This work was also funded in part by the National Center for Supercomputing Applications and the University of Illinois, under the auspices of the NCSA/UIUC faculty fellows program.

References

1. Wyse, N., Dubes, R., Jain, A.: A critical evaluation of intrinsic dimensionality algorithmsa critical evaluation of intrinsic dimensionality algorithms. In: Gelsema, E.S., Kanal, L.N. (eds.) Pattern recognition in practice, Amsterdam, pp. 415–425. Morgan Kauffman Publishers, Inc., San Francisco (1980)
2. Kim, K.A., Oh, S.Y., Choi, H.C.: Facial feature extraction using pca and wavelet multi-resolution images. In: Sixth IEEE International Conference on Automatic Face and Gesture Recognition, p. 439. IEEE Computer Society, Los Alamitos (2004)
3. Podolak, I.T.: Facial component extraction and face recognition with support vector machines. In: FGR 2002: Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition, Washington, DC, USA, p. 83. IEEE Computer Society, Los Alamitos (2002)
4. Pei, M., Goodman, E.D., Punch, W.F.: Pattern discovery from data using genetic algorithms. In: Proceeding of 1st Pacific-Asia Conference Knowledge Discovery & Data Mining, PAKDD 1997 (1997)

5. Liu, H., Motoda, H.: Feature extraction, construction and selection: a data mining perspective. SECS, vol. 453. Kluwer Academic, Boston (1998)
6. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182 (2003)
7. Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L. (eds.): Feature Extraction, Foundations and Applications. Springer, Heidelberg (2006)
8. Tackett, W.A.: Genetic programming for feature discovery and image discrimination. In: Proceedings of the 5th International Conference on Genetic Algorithms, pp. 303–311. Morgan Kaufmann Publishers Inc., San Francisco (1993)
9. Sherrah, J.R., Bogner, R.E., Bouzerdoum, A.: The evolutionary pre-processor: Automatic feature extraction for supervised classification using genetic programming. In: Proc. 2nd International Conference on Genetic Programming (GP 1997), pp. 304–312. Morgan Kaufmann, San Francisco (1997)
10. Kotani, M., Ozawa, S., Nakai, M., Akazawa, K.: Emergence of feature extraction function using genetic programming. In: KES, pp. 149–152 (1999)
11. Bot, M.C.J.: Feature extraction for the k-nearest neighbour classifier with genetic programming. In: Miller, J., Tomassini, M., Lanzi, P.L., Ryan, C., Tetamanzi, A.G.B., Langdon, W.B. (eds.) EuroGP 2001. LNCS, vol. 2038, pp. 256–267. Springer, Heidelberg (2001)
12. Zhang, Y., Rockett, P.I.: A generic optimal feature extraction method using multiobjective genetic programming. Technical Report VIE 2006/001, Department of Electronic and Electrical Engineering, University of Sheffield, UK (2006)
13. Guo, H., Nandi, A.K.: Breast cancer diagnosis using genetic programming generated feature. Pattern Recognition 39(5), 980–987 (2006)
14. Zhang, Y., Rockett, P.I.: A generic multi-dimensional feature extraction method using multiobjective genetic programming. Evolutionary Computation 17(1), 89–115 (2009)
15. Harris, C.: An investigation into the Application of Genetic Programming techniques to Signal Analysis and Feature Detection,September. University College, London (September 26, 1997)
16. Smith, M.G., Bull, L.: Genetic programming with a genetic algorithm for feature construction and selection. Genetic Programming and Evolvable Machines 6(3), 265–281 (2005)
17. Wang, K., Zhou, S., Fu, C.A., Yu, J.X., Jeffrey, F., Yu, X.: Mining changes of classification by correspondence tracing. In: Proceedings of the 2003 SIAM International Conference on Data Mining, SDM 2003 (2003)
18. Yang, Y., Wu, X., Zhu, X.: Conceptual equivalence for contrast mining in classification learning. Data & Knowledge Engineering 67(3), 413–429 (2008)
19. Cieslak, D.A., Chawla, N.V.: A framework for monitoring classifiers' performance: when and why failure occurs? Knowledge and Information Systems 18(1), 83–108 (2009)
20. Koza, J.: Genetic Programming: On the Programming of Computers by Means of Natural Selection. The MIT Press, Cambridge (1992)
21. AmericanCancerSociety: How many men get prostate cancer? `http://www.cancer.org/docroot/CRI/content/CRI_2_2_1X_How_many_men_get_prostate_cancer_36.asp`
22. Fernandez, D.C., Bhargava, R., Hewitt, S.M., Levin, I.W.: Infrared spectroscopic imaging for histopathologic recognition. Nature Biotechnology 23(4), 469–474 (2005)

23. Levin, I.W., Bhargava, R.: Fourier transform infrared vibrational spectroscopic imaging: integrating microscopy and molecular recognition. Annual Review of Physical Chemistry 56, 429–474 (2005)
24. Llorà, X., Reddy, R., Matesic, B., Bhargava, R.: Towards better than human capability in diagnosing prostate cancer using infrared spectroscopic imaging. In: Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation GECCO 2007, pp. 2098–2105. ACM, New York (2007)
25. Llorà, X., Priya, A., Bhargava, R.: Observer-invariant histopathology using genetics-based machine learning. Natural Computing: An International Journal 8(1), 101–120 (2009)
26. Quinlan, J.R.: C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco (1993)
27. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7, 1–30 (2006)
28. García, S., Herrera, F.: An extension on 'statistical comparisons of classifiers over multiple data sets' for all pairwise comparisons. Journal of Machine Learning Research 9, 2677–2694 (2008)
29. García, S., Fernández, A., Luengo, J., Herrera, F.: A study of statistical techniques and performance measures for genetics-based machine learning: Accuracy and interpretability. Soft Computing 13(10), 959–977 (2009)
30. García, S., Fernández, A., Luengo, J., Herrera, F.: Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. Information Sciences 180(10), 2044–2064 (2010)
31. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bulletin 1(6), 80–83 (1945)
32. Sheskin, D.J.: Handbook of Parametric and Nonparametric Statistical Procedures, 4th edn. Chapman & Hall/CRC (2007)

Label-free characterization of cancer-activated fibroblasts using infrared spectroscopic imaging

S. E. Holton,[†‡] M. J. Walsh,[‡] A. Kajdacsy-Balla,[§] and R. Bhargava, [†‡¶*]

Condensed Running Title:

Label-free imaging of fibroblasts

Keywords:

α-SMA, myofibroblast, tumor progression, vibrational spectroscopy, three-dimensional cell culture

ABSTRACT

Glandular tumors arising in epithelial cells comprise the majority of solid human cancers. Glands are supported by stroma, which is activated in the proximity of a tumor. Activated stroma is often characterized by the molecular expression of α-smooth muscle actin (SMA) within fibroblasts. The precise spatial and temporal evolution of chemical changes in fibroblasts upon epithelial tumor signaling, however, is poorly understood. Here we report a label-free method to characterize fibroblast changes using Fourier transform infrared (FT-IR) spectroscopic imaging by comparing spectra to α-SMA expression in primary normal human fibroblasts. The fibroblast activation process was recorded by spectroscopic imaging using increasingly tissue-like conditions – (a) simulation using the growth factor TGFβ1, (b) co-culture with MCF-7 human breast cancerous epithelial cells in Transwell co-culture and, (c) with MCF-7 in three-dimensional cell culture. Spectral signatures of stromal transformation were finally compared to normal and malignant human breast tissue biopsies. Results indicate that temporally complex spectral changes are observed, providing a richer assessment than simple molecular imaging based on α-SMA expression. Some changes are conserved across culture conditions and in human tissue, providing a label-free method to monitor stromal transformations.

[*] Author to whom correspondence should be addressed: rxb@illinois.edu

[†] Department of Bioengineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[‡] Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[§] Department of Pathology, University of Illinois at Chicago, Chicago, IL 60612, USA

[¶] Micro and Nanotechnology Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

INTRODUCTION

The stroma is known to play a crucial role in epithelial cancer progression in a variety of tissues (1-4). The stroma has also been suggested as an alternative and potentially more effective therapeutic target because the vast heterogeneity in the genomic and histological makeup of epithelial tumors makes individualized treatment expensive and unreliable (5). Methods to characterize the stroma, hence, and its transformations in epithelial tumor progression are imperative. One hallmark of a cancer-associated stroma, for example, is the fibroblast-to-myofibroblast cellular transformation (6). This phenotypic change is characterized by the expression of α-smooth muscle actin (α-SMA), a cytoplasmic protein that increases the cell's contractility and leads to the stiffening of the tumor microenvironment (7). The fibroblast-to-myofibroblast transformation has been observed within tumor-adjacent stroma in human tissues (8-10). A similar response can be induced by exposing fibroblasts to elevated levels of transforming growth factor-β1 (TGF-β1) in cell culture (11). Because of the readily-observable transition and its effect on physical properties of the tissue, stromal myofibroblasts have been a focus of research and are important markers in glandular cancers such as breast cancer (10,11). Immunohistochemistry (IHC) is the gold standard for visualizing α-SMA expression in clinical samples but using antibody-based techniques is time-consuming, costly and quantifying protein expression is difficult (12). The stromal response, further, is likely more complex than characterized by this single marker. Though advances in immunofluorescence have made considerable progress (13), only a few known proteins can simultaneously be detected. Even this capability may not be sufficient to catalogue the varied cytopathic effects of a multifactorial disease like cancer. Alternative techniques to directly measure cellular transformations in a consistent, quantitative and multiplexed manner are needed.

As an alternative to molecular imaging, label-free chemical imaging approaches have recently provided reliable correlations between histopathologic status and spectral markers (14-16). Fourier transform infrared (FT-IR) spectroscopic imaging, in particular, has been used extensively to study biochemical changes within cells as well as differences between cell lines (17-19). Molecular expression in simple breast cell cultures, too, has been correlated to spectral properties in both IR (20) and Raman spectroscopy (21). These studies have focused on epithelial cells. The fibroblast response to epithelial transformations has not been studied *in vitro* using spectroscopic imaging techniques. Here, we describe a method for characterizing and analyzing the fibroblast to myofibroblast differentiation. We specifically seek to examine the correlation between the current gold standard antibody marker and the spectroscopic signature of transformation. We examine transformation in primary normal dermal fibroblasts activated with TGFβ1, in co-cultures of primary fibroblasts with tumorigenic breast epithelial cells (MCF-7) and human tissues. While co-culture models provide a tissue-like environment, TGFβ1 activation is used as a positive control because it is commonly used in research (7). This comprehensive examination of cells grown in two-dimensional (2D) and three-dimensional (3D) cell cultures as well as human breast tissue will ensure wide research and clinical relevance. Finally, the presence of other cell types is a potentially confounding analytical factor and it is not obvious that spectral correlations will hold for mixtures of cell types. Hence, this study is important from the perspective of clinical cancer progression, for research in correlating labeled and label-free approaches as well as in the analysis of samples that present a complex bioanalytical background.

METHODS

Experimental Design

Cell Culture Models

To observe the effects of cancerous breast epithelium on the surrounding tissue stroma, two co-culture methods were utilized. First, the Transwell co-culture (Figure 1A) allows for two cell types to communicate via soluble growth factors that diffuse into a shared medium (22). Second, a 3D cellular co-culture model (Figure 1B) was developed. The 3D model consists of cells embedded in a type I collagen hydrogel. While both the systems are essentially mediated by soluble growth factors, cells adhere to a solid substrate in the 2D model and have a different geometry and physical microenvironment. While 2D monolayer cultures are the staple of cell biology, 3D cultures have recently been shown to be a more realistic representation of biological phenomena that occur in tissue (23-28). Hence, an analysis of both systems using our approach serves to ensure that the developed method is robust and relevant to different communities of researchers. In addition to these co-cultures, we sought to demonstrate that the developed methods are also valid for 2D and 3D cultures of single cell types. Therefore, we stimulated single cell type cultures with TGFβ1 to validate the observed activation.

For 2D cultures, primary normal human dermal fibroblasts (NDF) were grown on MirrIR slides, which allowed for both FT-IR transflectance and immunofluorescence imaging. The fibroblasts were co-cultured with cancerous breast epithelial cells, MCF-7, or stimulated with TGFβ1. MCF-7 cells were derived from a human breast tumor that had metastasized to the lung, but maintain a cancer stem cell-like phenotype in culture (29). They are less aggressive when injected into nude mice compared with other human breast cancer cell lines and, hence, were used in this study as a model of an ‘early’ cancerous source. Samples were removed from the culture at specific time points (0h, 6h, 12h, 24h) and fixed. Half the number of samples were analyzed using immunofluorescence to detect α-SMA. The other half were spectroscopically imaged. For 3D cultures, samples were prepared as separate layers, with one cell type (NDF, MCF-7) per layer in a type I collagen matrix. The layers are co-cultured for a determined length of time, and then separated with forceps. There was no observed cell migration within the time intervals of this experiment, determined by cell type specific expression of cytokeratins for epithelial cells and vimentin for fibroblasts (Data not shown). Briefly, the layers were separated, stained using standard IHC methods, and subsequently imaged using a Zeiss Axiovert 200M. A similar 3D model has been demonstrated previously to study skin cancer (29). While the experimental methods of both this study and the engineered skin model are capable of studying epithelial-fibroblast interactions, the pre-defined geometry here allows observations of molecular changes without morphology-associated effects or changing molecular concentration of a growing tumor that may confound the temporal profile.

Cell Culture

Cell lines and Use. Normal adult primary dermal fibroblasts (NHDF, Lonza, #CC-2511) were maintained in Fibroblast Basal Medium supplemented with 0.1% hFGF-B, 0.1% Insulin, 0.1% gentamicin/amphotericin-B, and 2% fetal bovine serum (FBS) (FGM-2 Fibroblast Growth Medium-2 Bullet Kit, Lonza, #CC3132). They were used at passage 8-10 to avoid problems associated with senescence in primary cell lines. The fibroblasts were subcultured according to protocols detailed on the Lonza website, and their ReagentPack (Trypsin/EDTA, Trypsin Neutralizing Solution, HEPES Buffered Saline Solution, #CC-5034) was used exclusively with this cell type. For serum-free medium, the media was prepared the same manner, except that FBS was omitted. MCF-7 (ATCC) cells were maintained in Dulbecco's Modified Eagle's Essential Medium (Invitrogen) supplemented with 10% FBS (Sigma) and 1% PenStrep (Sigma). They were subcultured according to ATCC protocols every 3 days at 70% confluency.

2D cell culture. Fibroblasts were grown on sterilized MirrIR slides (Kevley Technologies, Chesterland, Ohio, USA). They were seeded at approximately 60% confluency and grown for 24 hours before being switched to serum-free medium (FGM-2, Lonza, with additives but not FBS). The samples were grown in serum-free medium for 24 hours before co-culture. MCF-7 were grown on transwell inserts (Corning, 0.1μm pore size) in normal growth medium for 24 hours and then switched to serum-free medium for an additional 24 hours before co-culture.

Transwell co-culture. The Transwell co-culture system is useful for spectroscopy, because any IR substrate can be used in the lower chamber of the culture dish (Transwell inserts, 0.1 μm pore, PES, Corning Incorporated, Corning, NY, USA). Here, pieces of MirrIR Low-E slides were sterilized once with 10% bleach followed by 70% ethanol and left to dry in a sterile biosafety cabinet before use. Immunofluorescence staining was also performed using the MirrIR slides, and there were no detrimental effects on the coated glass surface. Using the MirrIR slides for both immunofluorescence and FT-IR measurements ensured that there was no substrate-specific factor that could have induced α-SMA expression independent of soluble growth factors. After 0h, 6h, 12h, and 24h of co-culture, each MirrIR slide was rinsed with sterile 1X PBS before fixation in 4% paraformaldehyde for 1 hour at 4˚C. After fixation, paraformaldehyde was neutralized with 0.1 M glycine for 10 minutes. Subsequently, the samples were divided: for each time point, two samples were prepared for FT-IR imaging and two were prepared for immunofluorescence staining. The samples for FT-IR imaging were rinsed with de-ionized water and left to dry prior to imaging.

3D cell culture. Cells were maintained as previously described in two-dimensional culture before being suspended in collagen hydrogels (Type I derived from rat tail, BD Biosciences). All reagents were kept on ice before plating because collagen solution will gel slightly at room temperature. In a conical tube on ice, collagen stock solution was diluted to 2 mg/mL with sterile 10X PBS. Cells were trypsinized, centrifuged at 1000 rpm for 3 minutes, and resuspended in growth medium. After counting, cells were suspended in the collagen solution at a cell density of 420 cells/mL for NDF and 1.9x10^4 cells/mL MCF-7. A much lower cell density of fibroblasts was used compared with epithelial cells because of the tendency of fibroblasts to collapse the hydrogel at high cell density and after activation and the similarity to fibroblast density in real tissue. Finally, 1N NaOH was added at 0.023 mL per 1 mL of collagen stock solution to neutralize the acetic acid and allow the collagen to gel. To prepare samples, 200 μL of collagen and cell suspension was added to each well of a 48-well tissue culture plate. The plates were left at 4˚C for 90 minutes to slow down the polymerization of collagen to provide a more uniform fiber orientation and width (31). The samples were then placed in a humidified incubator at 37˚C for 30 minutes to polymerize the collagen with cells embedded within. After the samples had gelled, growth medium was added. The cells were allowed to grow for 24 hours before being changed to serum-free medium to avoid any confounding effects of growth factors present in FBS. After 48 hours in serum-free medium, the co-culture layers were stacked together and 1.5 ng/mL TGFβ1 (Transforming Growth Factor-β1 from human platelets, ≥ 97%, Sigma, #T1654) in serum-free medium was added to the appropriate fibroblast samples as a positive control. Fresh serum-free medium was added to the co-cultured samples. After 0, 6, 12, and 24 hours of co-culture, the fibroblast layer was fixed in 4% paraformaldehyde overnight before processing for immunofluorescence or FT-IR imaging.

Three-dimensional culture sample preparation. 3D culture samples were paraffin embedded and sectioned prior to imaging. First, the paraformaldehyde was gently aspirated from the samples and then the gels were dehydrated by serial ethanol dehydration. The samples were put in 50%, 70%, 80%, and 95% ethanol for 45 minutes each followed by three 45 minute incubations in 100% ethanol. The samples were then soaked in xylenes for three 45 minute periods. Finally, the samples were placed in paraffin in a 60˚C oven for two 1 hour periods and one 12 hour period. The samples were mounted in paraffin blocks and sectioned at 5 μm onto MirrIR slides for FT-IR imaging. Samples were de-paraffinized in hexanes for 24 hours before imaging. For each set of experiments, samples were prepared in duplicate and the experiment was replicated independently to show reproducibility of both biological results and absorbance spectra.

Immunofluorescence Staining. For immunofluorescence staining, samples were permeabilized in 0.2% TX-100 for 15 minutes. After washing three times with PBS the samples were blocked with a 1wt% BSA in PBS/T for 1.5 hours. After three washes with PBS/T, the samples were incubated with primary antibody (Mouse anti-human α-SMA, Dako, 1:100 dilution) overnight at 4˚C. The samples were washed again and incubated with secondary antibody (Goat anti-Mouse IgG-FITC conjugated, abcam, 1:80 dilution) for one hour. The samples were mounted with UltraCruz Mounting Medium for Fluorescence with DAPI (Santa Cruz Biotechnology, Cat # sc-24941) and imaged using a Zeiss Axiovert 200M fluorescence microscope. For three-dimensional samples, confocal imaging was done using a Leica SP2 laser scanning confocal microscope.

Immunohistochemistry: Tissue Biopsies. A tissue microarray (TMA) of 96 1.5mm human breast tissue cores comprising of normal, epithelial hyperplasia, in-situ, benign tumors and malignant cancer tissues was obtained (US Biomax, Inc. USA. #BR961). Four serial sections were acquired from the TMA block, one 5μm thick tissue section was placed on a BaF_2 substrate for FT-IR analyses and three 5μm thick tissues sections were placed on standard glass slides for IHC and hematoxylin and eosin (H&E) staining. IHC staining was performed for vimentin and α-SMA. H&E staining was used for tissue visualization. Staining was performed using a Ventana Benchmark XT Automated Slide Preparation system (Ventana Medical Systems, Inc.) and Ventana clinical protocols and reagents (XT UltraView DAB protocol, Ventana, Tucson, AZ).

FT-IR spectroscopic imaging. FT-IR spectroscopic imaging data were recorded using a Perkin Elmer Spotlight 400 imaging system. For all cellular samples, both confluent and sparse regions of the sample were imaged in a transflection mode and data from 4000 cm^{-1} to 750 cm^{-1} were saved. A spectral resolution of 8 cm^{-1} was set with 32 scans per pixel averaged to provide higher signal to noise ratio data. An interferometer speed of 1.0 cm/s was used while a pixel size of 6.25 x 6.25 μm was used for detection using a MCT linear array. A spectral background was collected on the MirrIR slide using the same parameters but with 120 scans per pixel. Atmospheric correction was performed on the Spotlight instrument, and the files were exported into ENVI-IDL. Images were baseline corrected and only those pixels with an absorbance greater than 0.015 a.u. for the peak absorbance at 1656 cm^{-1} (Amide I) were used for further analysis. Spectra were normalized to Amide I to account for variances in cell density.

For tissue microarray data, absorbance was stronger and we sought to maintain compatibility with earlier studies on the parameters used. Sample data were collected using the same scanning parameters as for cell culture samples, except a 4 cm^{-1} resolution with 2 scans per pixel and a mirror speed of 2.2 cm/s were used. A background was acquired at these parameters with 120 scans averaged. A threshold absorbance of 0.03 a.u. for the 1656 cm^{-1} (Amide I) absorbance peak was employed to determine pixels to be included in the analysis. Regions of Interest (ROIs) were manually marked on the absorbance images corresponding to regions of either fibroblast or myofibroblast cells. Cell-type assignations were made based on the IHC staining of the serial tissue sections; fibroblasts stained positive for vimentin and negative for α-SMA, and myofibroblasts stained positive for both vimentin and α-SMA. Over 40,000 pixels corresponding to fibroblasts and over 150,000 pixels corresponding to myofibroblasts were identified. From these identified pixels, average spectra were obtained for fibroblast and myofibroblast classes.

RESULTS AND DISCUSSION

The well-characterized fibroblast activation pathway serves as a model system to benchmark spectral (chemical) changes that accompany the phenotypic transformation. The transwell co-culture system was used first to determine whether co-culturing normal primary fibroblasts with tumorigenic breast epithelial cells could result in an activated phenotype, as shown previously in fibroblasts isolated from stroma surrounding a tumor *in vivo* (7) as well as after induction by TGFβ1 *in vitro* (9). In our co-culture with MCF-7 cells, phenotypic changes were induced in the primary dermal fibroblasts within 6 hours to the same extent as treatment with 1.5 ng/mL TGFβ1 (Figure 2). The experiment was repeated for both cases over a time course of 24 hours, with timepoints being taken at 0 (no co-culture), 6, 12, and 24 hours to observe any potential evolution of this marker over time. From immunofluorescence imaging results, there was no visible change in the number of cells expressing α-SMA expression over time. No digitally-assisted methods were used in order to compare intensity levels as quantitative intensity analysis is difficult due to non-specific fluorescence and photobleaching. Both stimulation with TGFβ1 and co-culture with MCF-7 activated fibroblasts within 6 hours.

We hypothesized that examining the temporal evolution of IR absorption spectra would yield more information about fibroblast activation than the “on-off” information derived from immunofluorescence expression of a single biomarker. Spectra measured from fibroblasts are shown in Figure 3. Changes were primarily seen in the biomolecular fingerprint region (1800-950 cm^{-1}) and also in the C-H stretching region (3000-2875 cm^{-1}). In the fingerprint region, larger changes were seen in peaks at 1080 cm^{-1} and 1224 cm^{-1} (Figure 3, *top*). These are the asymmetric and symmetric vibrational modes of the phosphate bond, indicative of changes in nucleic acids. There is an increase in the 1080 cm^{-1} peak, which is usually associated with the symmetric phosphate stretching of DNA. These spectra are averaged to cell density, and the cells were serum-starved before the experiment began, and so there should have been little cellular proliferation over the 24 hour time course. Serum-starving the cells prior to co-culture arrests them at G0/G1, and this should minimize spectral differences in 1080 cm^{-1} that can be attributed to cells being in different phases of the cell cycle (32-34). The increase at 1080 cm^{-1} indicates that unless there is an increased amount of DNA present in the cells, this assignment of this peak to the phosphate bond of DNA alone may be uncertain. If we account for the total amount of genetic material present within the cell (RNA, DNA, and associated proteins), this could provide some explanations for the increases in absorbance at this peak. The spectral changes seen could be due to an increase in RNA, changes in chromatin three-dimensional configuration, chromatin sequestration, or an increase in the size of the nucleus. Recently reported by Whelton *et al*, changes in the 1080 cm^{-1} peak are attributed to a transition between native B- and A-like forms of DNA upon dehydration of intact cells. We do not anticipate that changes seen in these experiments are due to this transition because all samples were fixed and dried completely prior to spectroscopic imaging (35).

Between the two treatments (MCF-7 co-culture and TGFβ1 stimulation), there was similar molecular expression of α-SMA, but differences in absorption at 1080 cm^{-1}. In the TGFβ1 stimulated samples, the 6- and 12- hour samples had an increase in absorption at 1080 cm^{-1} compared with the control, but after 24 hours the level had fallen back to the control value. Interestingly, at 1224 cm^{-1}, the 6- and 24-hour time points were elevated while the 12-hour sample had lower absorption than the control. This discrepancy could be the result of the cells only being stimulated with TGFβ1 once at the beginning of the experiment. Thus, the 6 hour sample would have a sustained level of TGFβ1 in the medium before the cells were fixed, whereas in the 24 hour sample the concentration of TGFβ1 present in the medium has decreased because it has already been metabolized by the cells. However, in the samples that were co-cultured with MCF-7 cells, there was a uniform level of growth factors secreted by the epithelial cells into the shared medium throughout the time course of the experiment. Therefore, we believe that the absorbance at 1224 cm^{-1} may be used as a marker for a sustained fibroblast response to molecular signals released by a malignant epithelium.

In the C-H stretching region, changes were seen in peaks at 2850 cm^{-1}, 2930 cm^{-1}, and 2960 cm^{-1}. This region of the spectrum is correlated with proteins and also the carbonyl chains of fatty acids (31). With increasing lengths of time after TGFβ1 stimulation, there was a gradual increase in peak height across all peaks in this region (Figure 3B, *bottom*). In contrast, co-culture with MCF-7 cells yielded a fibroblast response that was more defined, with a very rapid increase in peak height at 2930 cm^{-1} after just 6 hours in comparison with the control (Figure 3A, *bottom*). Although immunofluorescence results show α-SMA expression in samples stimulated with TGFβ1 or co-cultured with MCF-7 cells, there were differences in absorption spectra between the two sets of samples, permitting a more in-depth biochemical analysis of cellular activation. The reasons for the difference in kinetics of activation likely stem from the co-culture providing a host of molecules in the activation pathway via paracrine signaling. While the mechanisms of the two activations are likely different, this would not be apparent from a single marker. It is also interesting to contrast the ability of spectroscopy to measure transient behavior, which is lacking in the expression of α-SMA. Vibrational spectroscopy, of course, does not provide specific protein expression levels in cells. As a general strategy for comprehensive biomolecular analysis, hence, the spectroscopic data can be used to inform the search for appropriate molecular markers by providing the temporal evolution profiles. Further, there may be cellular and sub-cellular spectral heterogeneity across the sample upon stimulation. These have been examined elsewhere (36).

In contrast to the 2D transwell co-culture, culturing cells in a 3D geometry provides an environment that is closer to cellular chemistries *in vivo*. Cells are known to express surface receptors more faithfully in three-dimensional culture and are also more likely to differentiate in response to external stimuli (37-39). In the 3D co-culture system described here, a single cell type and collagen scaffold was used fabricate a cylindrical-shaped "layer" (Figure 1B). Layers containing different cell types were prepared separately and subsequently stacked on top of each other. This technique allowed for cells to be co-cultured by simple stacking. Since the layers are only weakly adherent, they could subsequently be mechanically re-separated for analysis. As previously used in 2D cell culture, immunofluorescence staining for α-SMA was used to probe for the presence of myofibroblasts in 3D using confocal microscopy (Figure 2C). The immunofluorescence results remained consistent with the transwell co-culture results; exposure to MCF-7 cells activated fibroblasts along the same time course as TGFβ1 exposure. Another use for three-dimensional cell culture in this setup is that the collagen peaks (1283 cm^{-1}, 1236 cm^{-1}, and 1204 cm^{-1}) can be used for IR spectral analysis—either as control or for examination of microenvironmental changes associated with a growing tumor. These collagen peaks are diagnostically useful when looking at whole tissue sections (15), and there is evidence showing that changes in collagen spectra can be detected within a certain distance from a tumor (30,40), which is clinically relevant for cancer pathology. Hence, we examined the same in the 3D co-culture model (Figure 4A).

The only major observation in our study was an overall increase in the absorption of the collagen peaks after co-culture with MCF-7 cells over time. This could be a result of fibroblasts locally depositing collagen upon exposure to MCF-7 stimuli. Myofibroblasts play an important role in tissue maintenance, providing a wound healing-type response by depositing more collagen in the surrounding extracellular matrix (41). TGFβ1 also stimulates fibroblasts to deposit collagen via the Smad pathway, which aids in the transcription of the α2(I) procollagen gene, COL1A2 (42). It is suggested that the fibroblasts present in collagen-dense keloid scars are more susceptible to TGFβ1 (43). Further, the extracellular matrix can act as a control mechanism for the involvement of TGFβ1 in collagen biosynthesis (44). The other possibility is that upon fibroblast activation, the stiffening of the cells themselves results in the contraction of the surrounding gel, making local regions appear more collagen-dense in the absorption spectra. However, no detectable gel contraction was observed upon visual inspection during the timecourse of this experiment, likely due to the low cell density of fibroblasts embedded within the collagen matrix. For these reasons, we believe that spectral changes seen in this model are indicative of collagen remodeling by cancer-activated fibroblasts.

Consistent with the 2D culture results, changes were seen in the 1080 cm^{-1} peak in the 3D culture model. There was an increase in this peak initially, however after 24 hours this peak has diminished. The 'ebb and flow' of this nucleic acid signature, even in the environment of persistent epithelial cues, suggests that fibroblasts' molecular expressions settle into a new equilibrium upon an initial exposure to transforming stimuli. This observation is also consistent with changes seen in the Transwell co-culture (Figure 3, *bottom*). There is no change in RNA levels (1224 cm^{-1}) seen in this model compared with the Transwell-culture model. This could be a result of diminished cytoplasmic material to record data from as cells appear smaller in the 3D matrix and thus the cytoplasm is much smaller compared to the nucleus of the cells. Peaks in the C-H stretch region, as with other cultures, may correlate with changes in the phospholipid membrane or protein synthesis after fibroblast activation. Either explanation is plausible considering the physiologic changes that occur during the fibroblast to myofibroblast phenotypic change. However, in the 12- and 24- hour time points, there is a significant decrease in absorption in this region compared with the control. In general, absorbance in this area was low compared with samples cultured in monolayers due to cells in 3D cultures being sparsely populated and of thinner shape than 2D cultures. Thus, their accurate monitoring is much more challenging than 2D monolayers. In the C-H stretching region, biochemical changes are dominated by events occurring in the cytoplasm of cells (36) as were the changes at 1224 cm^{-1}. In the 3D culture, as with tissues, it is productive to examine changes due to cellular secretion of growth factors in the surrounding extracellular matrix. Examining changes within the cells themselves requires a subcellular localization of signals that is not achieved here.

To translate the understanding from these studies, clinical breast tissue samples were examined by IR imaging and immunohistochemical staining, including for vimentin and α-SMA. Vimentin (Figure 5B) will stain for fibroblasts and other mesoderm-derived tissues. In contrast, α-SMA (Figure 5C) is a protein found in myofibroblasts, myoepithelium that lines each gland, and smooth muscle cells which surround blood vessels. Hence, we were able to differentiate between normal and activated fibroblasts by comparing the localization of these two markers in adjacent sections of tissue. In the clinical breast tissue samples, vimentin (in brown) is primarily seen in the stroma between glands (Figure 5B). However, only fibroblasts nearest the cancerous epithelium express α-SMA (Figure 5C). This is a cancer-associated signature and is diagnostically relevant. In order to provide a critical comparator to the work performed in our monolayer and three-dimensional co-culture models, we examined spectral differences between activated and resting-state fibroblasts in these clinical samples. IR spectroscopic imaging was performed on an entire TMA. Based on the staining of adjacent sections, pixels were manually marked and classified as 'fibroblast' or 'myofibroblast'. The pixels for each class were averaged and these average spectra across the TMA were examined, as shown in Figure 6. Spectra were compared with the 3D results, because this model should be biologically closest to clinical samples. However, upon examination, the results between the two models are inconsistent. Although the collagen peaks (1300-1050 cm^{-1}) and C-H stretching region (3000–2800 cm^{-1}), are consistent in *shape* between the three-dimensional culture model and the tissue sample, myofibroblasts from the clinical samples show lower absorbance in the biomolecular fingerprint region (Figure 6A). Spectra from the 3D cultures were ‘pure’, consisting only of normal or activated fibroblasts and type I collagen. Using clinical samples is invaluable, but leads to more variables that become increasingly difficult to control. For example, there is a large degree of variance between patients, even for the noncancerous biopsies (unpublished data, M.J. Walsh). Another interesting avenue is whether immuhistochemical stains, used here as the gold standard for comparison, are truly as reliable in clinical samples as in cell culture studies.

Across the three systems described in this manuscript, activated fibroblasts display spectral changes in the mid-IR regions associated with nucleic acids (1080 cm^{-1}, 1224 cm^{-1}) and C-H stretching modes (2850 cm^{-1}, 2930 cm^{-1}, 2960 cm^{-1}). Although 2D and 3D co-culture models were mostly consistent, we found some discrepancies between the *in vitro* and the clinical specimens. Studying this transition with FT-IR spectroscopy under controlled cell culture conditions yields important information about the potential kinetics of paracrine signaling between epithelial cells and fibroblasts. Investigating the C-H stretching region of fibroblasts also results in overall increased absorbance at all three peaks across all scenarios, including human breast tissue biopsies (Figure 6B). The nature of fibroblast activation involves a cellular phenotypic change, where cytoplasmic proteins are produced, and the shape of the cell undergoes a transformation. These biological phenomena can be correlated with the increased absorbance in peaks associated with C-H stretching as a marker for a cancer-activated stromal profile. Because FT-IR spectroscopic imaging can be used to study the distribution of chemical changes across the area of a sample, this understanding can be applied to detect early stromal activation in noncancerous areas of a biopsy or tissue resection independent of the expression of a biomarker. This same technique could be expanded to different biological problems, such as testing the effects of drug delivery on distal tissues. By correlating these biological phenomena observed in cell culture with chemical signatures, label-free imaging in complex human tissues becomes elucidated.

FT-IR spectroscopy and imaging have been employed to measure a wide variety of biomolecular species, including nucleic acids, collagen, glycogen, proteins, and fatty acids. The complex mixtures of these molecules present in cells and tissues implies that IR spectroscopy is useful for determining global biochemical changes in classes of these materials and is sensitive to the metabolic (45) and local physiologic state of the tissue. In this study, we demonstrate that the method extracts more detailed changes compared to conventional immunofluorescence. The correlations of these changes with mechanistic molecular transitions in the cell can now be established. This next step will link many events in the transformation a simple, label-free measurement. Since IR imaging data are a convolution of the underlying spectral and structural properties of the tissue (46) and the imaging setup (47-49) and optical properties (50-52), measurement of specific molecular alterations becomes very challenging. Nevertheless, we show that there is conservation of some changes in the fibroblast-to-myofibroblast transformation that translates across monolayer culture, three-dimensional culture, and human tissues. In summary, IR absorption imaging provides a label-free approach for integrated, first-pass approaches that can yield information about changes in the sample. Such information can provide a basis for studies by itself or an early indication of which biological assays to perform next and is especially critical for heterogeneous samples in which we need to determine where to perform further molecular analysis.

CONCLUSIONS

Normal adult human fibroblasts were examined in monolayer and three-dimensional cell cultures as well as formalin-fixed and paraffin embedded human tissue to correlate the expression of α-SMA using immunofluorescence techniques to chemical changes, as observed using FT-IR spectroscopic imaging. Spectral changes were observed predominantly in the C-H stretching region (3290 cm^{-1}) and phosphate bonds associated with nucleic acids (1224 cm^{-1} and 1080 cm^{-1}). In 3D co-cultures and human tissue biopsies, the microenvironmental changes were assessed by examining vibrational modes commonly associated with collagen (1283 cm^{-1}, 1236 cm^{-1}, and 1204 cm^{-1}). Fibroblasts activated *in vitro* via TGFβ1 stimulation or co-culture with breast cancer epithelial cells expressed α-SMA and were spectrally distinct from resting-state fibroblast controls. This was also true in the tissue biopsies. However, the spectra from cellular cultures were not entirely consistent with those from tissue, particularly in the phosphate peaks. Although the overall spectral characteristics are conserved between the 3D culture and biopsies, specific absorbance values were inconsistent. Furthermore, there is a spatial dependence of this expression based on the distance of the fibroblasts from the tumor ‘source’, determined by analysis of the collagen peaks and expression of α-SMA in tissue. By directly extracting spectral signatures of fibroblast activation, analysis can potentially provide new information, be conducted in a high-throughput manner and reduce variability, time, and costs. Finally, this work exhibits a novel use of IR spectroscopic imaging in examining stromal changes associated with tumor progression.

ACKNOWLEDGEMENTS

We acknowledge support for this project from Komen for the Cure (KG081426), Department of Defense Prostate Cancer Research Program (W81XWH-07-1-0242) and National Institutes of Health (R01CA138882). S.E.H. was funded at UIUC from NIH National Cancer Institute Alliance for Nanotechnology in Cancer ‘Midwest Cancer Nanotechnology Training Center’ Grant R25 CA154015A.

REFERENCES

1. Illmensee, K., and B. Mintz. 1976. Totipotency and normal differentiation of single teratocarcinoma cells cloned by injection into blastocysts. Proc Natl Acad Sci USA. 73:549-53.

2. Atula, S., R. Grenman, and S. Syrjänen. 1997. Fibroblasts can modulate the phenotype of malignant epithelial cells in vitro. Exp Cell Res. 235:180-187.

3. Weaver, V.M., O.W. Petersen, F. Wang, C.A. Larabell, P. Briand, C. Damsky, and M.J. Bissell. 1997. Reversion of the malignant phenotype of human breast cells in three-dimensional culture and in vivo by integrin blocking antibodies. J Cell Biol. 137:231-245.

4. Dolberg, D.S., and M.J. Bissell. 1984. Inability of rous sarcoma virus to cause sarcomas in the avian embryo. Nature. 309:552-556.

5. Ingber, D.E. 2008. Can cancer be reversed by engineering the tumor microenvironment? Semin Cancer Biol. 18:356-364.

6. Barcellos-Hoff, M., and D. Medina. 2005. New highlights on stroma-epithelial interactions in breast cancer. Breast Cancer Res. 7:33-36.

7. Rønnov-Jessen, L. 1996. Stromal reaction to invasive cancer: The cellular origin of the myofibroblast and implications for tumor development. The Breast Journal. 2: 320-339.

8. Bhowmick, N., Neilson, E., and Moses, H. 2004. Stromal fibroblasts in cancer initiation and progression. Nature, 432: 332-337.

9. Hawsawi, N.M., H. Ghebeh, et al. 2008. Breast carcinoma-associated fibroblasts and their counterparts display neoplastic-specific changes. Cancer Res. 68:2717-25.

10. Bauer, M., G. Su, C. Casper, et al. 2010. Heterogeneity of gene expression in stromal fibroblasts of human carcinomas of the breast. Oncogene. 29:1732-1740.

11. Rønnov-Jessen, L., and O. Petersen. 1995. The origin of myofibroblasts in breast cancer. Recapitulation of tumor environment in culture unravels diversity and implicates converted fibroblasts and recruited smooth muscle cells. J Clin Invest. 95:859-873.

12. Camp, R.L., G. G. Chung, and D. L. Rimm. 2002. Automated subcellular localization and quantification of protein expression in tissue microarrays. Nat Med. 8:1323-1328.

13. Giltnane, J.M., J.R. Murren, D.L. Rimm, and B.L. King. 2006. AQUA and FISH analysis of HER-2/neu expression and amplification in a small cell lung carcinoma tissue microarray. Histopathology. 49:161-169.

14. Diem, M., P. Griffiths, and J. Chalmers. 2008.Vibrational techniques in medical diagnostics. Wiley, New York.

15. Fernandez, D. C., R. Bhargava, S.M. Hewitt, and I.W. Levin. 2005. Infrared spectroscopic imaging for histopathologic recognition. Nat Biotechnol. 23:469-47

16. German, M.J., A. Hammiche, N. Ragavan,..., and F.L. Martin. 2006. Infrared spectroscopy with multivariate analysis potentially facilitates the segregation of different types of prostate cell. Biophys J. 90:3783-95.

17. Zhang, L., G.W. Small, A.S. Haka, L.H. Kidder, and E.N. Lewis. 2003. Classification of Fourier transform infrared microscopic imaging data of human breast cells by cluster analysis and artificial neural networks. Appl Spectrosc. 57:14-22.

18. Bogomolny, E., Huleihel, M., & Suproun, Y. 2007. Early spectral changes of cellular malignant transformation using Fourier transform infrared microspectroscopy. J Biomed Opt. 12: 024003-1-9.

19. Diem, M., C. Matthäus, T. Chemenko, M.M. Romeo, M. Miljković, B. Bird, J. Schubert, K. Papamarkakis, and N. Laver. 2009. Infrared and Raman Spectroscopy and Spectroscopic Imaging of Individual Cells. In Infrared and Raman spectroscopic imaging. R. Salzer and H.W. Siesler, editors. Wiley-VCH, Weinheim. 173-188.

20. Yang, W., X. Xiao, J. Tan, and Q. Cai. 2009. In situ evaluation of breast cancer cell growth with 3D ATR-FTIR spectroscopy. Vib Spectrosc. 49:64-67.

21. Hartsuiker, L., N. J. L. Zeijen, L. W. M. M. Terstappen, and C. Otto. 2010. A comparison of breast cancer tumor cells with varying expression of the Her2/neu receptor by Raman microspectroscopic imaging. Analyst, 135: 3220-3226.

22. Rozenchan, P., and D. Carraro. 2009. Reciprocal changes in gene expression profiles of cocultured breast epithelial cells and primary fibroblasts. Int J Cancer. 125:2767-2777.

23. Tibbitt, M., and K. Anseth. 2009. Hydrogels as extracellular matrix mimics for 3D cell culture. Biotechnol Bioeng. 103:655-663.

24. Yamada, K., and E. Cukierman. 2007. Modeling tissue morphogenesis and cancer in 3D. Cell. 130:601-610.

25. Martin, K., D. Patrick, M. Bissell, and M. Fournier. 2008. Prognostic breast cancer signature identified from 3D culture model accurately predicts clinical outcome across independent datasets. PLoS One. 3:e2994.

26. Liu, V.A., and S. Bhatia. 2002. Three-dimensional photopatterning of hydrogels containing living cells. Biomed Microdevices. 4:257-266.

27. Nelson, C.M., J. L. Inman, and M. J. Bissell. 2008. Three-dimensional lithography defined organotypic tissue arrays for quantitative analysis of morphogenesis and neoplastic progression. Nat Protoc. 3:674-678.

28. Pampaloni, F., E. Reynaud, and E. Stelzer. 2007. The third dimension bridges the gap between cell culture and live tissue. Nat Rev Mol Cell Biol. 8:839-845.

29. Mackenzie, I.C. 2006. Stem cell properties and epithelial malignancies. Eur J Cancer. 42:1204-1212.

30. Kong, R., R. Reddy, and R. Bhargava. 2010. Characterization of tumor progression in engineered tissue using infrared spectroscopic imaging. Analyst. 135:1569-1578.

31. Trier, S., K. Eliceiri, P. Keely, A. Friedl, and D. Beebe. 2009. Control of 3-dimensional collagen matrix polymerization for reproducible human mammary fibroblast cell culture in microfluidic devices. Biomaterials. 30:4833-4841.

32. Holman, H.N., M.C. Martin, E. A. Blakely, K. Bjornstad, and W.R. Mckinney. 2000. IR spectroscopic characteristics of cell cycle and cell death probed by synchrotron radiation based Fourier transform IR spectromicroscopy. Biopolymers. 57:329-335.

33. Hammiche, A., M.J. German, R. Hewitt, H.M. Pollock, and F.L. Martin. 2004. Monitoring cell cycle distributions in MCF-7 cells using near-field photothermal microspectroscopy. Biophys J. 88:3699-3706.

34. Flower, K.R., I. Khalifa, P. Bassan, D. Démoulin, E. Jackson, N.P. Lockyer, A.T. McGown, P. Miles, L. Vaccari, and P. Gardner. 2011. Synchrotron FTIR analysis of drug treated ovarian A2780 cells: an ability to differentiate cell response to different drugs? Analyst. 136:498-507.

35. Whelan, D.R., K.R. Bambery, P. Heraud, M.J. Tobin, M. Diem, D. McNaughton, and B.R. Wood. 2011. Monitoring the reversible B to A-like transition of DNA in eukaryotic cells using Fourier transform infrared spectroscopy. Nucleic Acids Res. doi:10.1093/nar/gkr175.

36. Holton, S.E., M.J. Walsh, and R. Bhargava. 2011. Subcellular localization of early biochemical transformations in cancer-activated fibroblasts using infrared spectroscopic imaging. Analyst, doi:10.1039/C1AN15112F

37. Menendez, J.A., and R. Lupu. 2007. Fatty acid synthase and the lipogenic phenotype in cancer pathogenesis. Nat Rev Cancer. 7:763-777.

38. Albrecht, D., G. Underhill, and T. Wassermann. 2006. Probing the role of multicellular organization in three-dimensional microenvironments. Nat Methods. 3:369-375.

39. Cukierman, E., R. Pankov, and K. Yamada. 2002. Cell interactions with three-dimensional matrices. Curr Opin Cell Biol. 14:633-639.

40. Dhimolea, E., M. Maffini, A. Soto, and C. Sonnenschein. 2010. The role of collagen reorganization on mammary epithelial morphogenesis in a 3D culture model. Biomaterials. 31:3622-3630.

41. Desmoulière, A., C. Guyot, and G. Gabbiani. 2004. The stroma reaction myofibroblast: a key player in the control of tumor cell behavior. Int J Dev Biol. 48:509-517.

42. Chen, S., W. Yuan, Y. Mori, A. Levenson, M. Trojanowska, and J. Varga. 1999. Stimulation of type I collagen transcription in human skin fibroblasts by TGF-β: Involvement of Smad 3. J Invest Dermatol. 112:49-57.

43. Bettinger, D. A., D.R. Yager, R.F. Diegelmann, and I.K. Cohen. 1996. The effect of TGF-β on keloid fibroblast proliferation and collagen synthesis. Plast Reconstr Surg. 98:827-833.

44. Streuli, C.H., C. Schmidhauser, M. Kobrin, M.J. Bissell, and R. Derynck. 1993. Extracellular matrix regulates expression of the TGF-β1 gene. J Cell Biol. 120:253-260.

45. Ellis, D. I., and R. Goodacre. 2006. Metabolic fingerprinting in disease diagnosis: biomedical applications of infrared and Raman spectroscopy. Analyst, 131: 875-885.

46. Bhargava, R. 2007. Towards a practical Fourier transform infrared chemical imaging protocol for cancer histopathology. Anal Bioanal Chem. 389:1155-1169.

47. Levin, I.W., and R. Bhargava. 2005. Fourier transform infrared vibrational spectroscopic imaging: Integrating microscopy and molecular recognition. Annu Rev Phys Chem. 56:429-74.

48. Diem, M. et al. 2004. A decade of vibrational micro-spectroscopy of human cells and tissue (1994)-2004). Analyst. 129:880-885.

49. Baker, M. J., E. Gazi, M. D. Brown, J.H. Shanks, N.W. Clarke, and P. Gardner. 2009. Investigating FTIR based histopathology for the diagnosis of prostate cancer. J Biophotonics. 2:104-113.

50. Davis, B.J., S. Carney, and R. Bhargava. 2010. Theory of infrared microspectroscopy for intact fibers. Anal Chem. doi:10.1021/ac102239b.

51. Tseng, S., A. Grant, and A.J. Durkin. 2008. In vivo determination of skin near-infrared optical properties using diffuse optical spectroscopy. J Biomed Opt. 13:0140161-7.

52. Cuccia, D.J., F. Bevilacqua, A.J. Durkin, F.R. Ayers, and B.J. Tromberg. 2009. Quantitation and mapping of tissue optical properties using modulated imaging. J Biomed Opt. 14:0240121-13.

Figure Legends

Figure 1. (A) Schematic of the trans-well co-culture system that allows the cells to communicate via soluble growth factors without contact. A filter with a 0.2 μm pore size is used at the bottom of the top basket. (B) Schematic of the three-dimensional (3D) co-culture setup, which is comprised of cells embedded in a type I collagen gel. No membrane separates the layers.

Figure 2. After six hours (6h) of stimulation with 1.5 ng/mL TGFβ1 (A) or 6h co-culture with MCF-7 cells (B), α-SMA is expressed in dermal fibroblasts. Confocal microscopy (C) was used to visualize α-SMA expression in fibroblasts for 3D systems. Scale bar represents 50 μm.

Figure 3. *Top* In fibroblasts co-cultured with MCF-7 (A), or after 1.5 ng/mL TGFβ1 stimulation (B), normal dermal fibroblasts exhibit changes primarily in the asymmetric and symmetric phosphate stretching bands, indicating bulk changes in the quantity of nucleic acids over time, normalized to 1656 cm^{-1} (Amide I). Fibroblasts activated through co-culture show sustained levels of nucleic acids over time, whereas levels wane in TGFβ1 activated fibroblasts.

Bottom Comparison of the C-H-stretching region for fibroblasts co-cultured with MCF-7 cells (A) and TGFβ1 stimulated fibroblasts (B). Peaks in the C-H stretching region of the spectrum (2960 cm^{-1}, 2932 cm^{-1}, and 2850 cm^{-1}) have a much higher absorbance in the 12- and 24- hour timepoints compared with control. This suggests an increase in cell metabolism through the presence of higher amounts of fatty acids. After 6 hours of TGFβ1 stimulation, fibroblasts show lower absorbance in this region compared with control and MCF-7 co-culture.

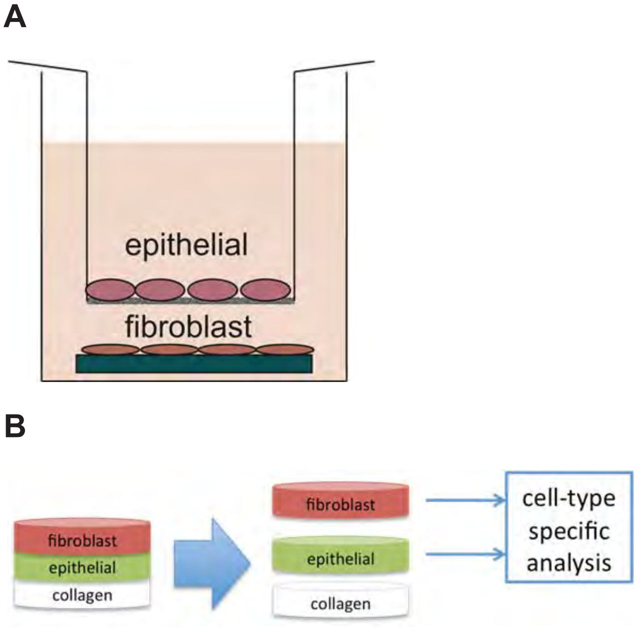
Figure 4. Characteristic absorbance peaks associated with collagen (1283 cm^{-1}, 1236 cm^{-1}, and 1204 cm^{-1}) are visible and elevated in fibroblasts after co-culture with MCF-7 (A). At 1080 cm^{-1} in both three-dimensional and two-dimensional culture (Figure 3B) the same cyclical phenomenon is shown. The C-H stretching region of the spectrum (B) is distinct from that of the transwell co-culture (Figure 4B) spectra.

Figure 5. Cancerous breast tissue biopsies demonstrate glandular and stromal regions to examine α-SMA expression proximal to cancerous epithelium. The morphological features are distinguished using hematoxylin and eosin staining (A). Fibroblasts are discerned by using IHC staining for vimentin (B). IHC staining for α-SMA (C) is positive for activated fibroblasts (myofibroblasts), myoepithelium (found lining the gland), and smooth muscle (found around blood vessels). α-SMA positive fibroblasts are located adjacent to the cancerous epithelium, but distal fibroblasts are negative for this protein. Each tissue core, part of a tissue microarray (TMA), is 0.5 mm in diameter.

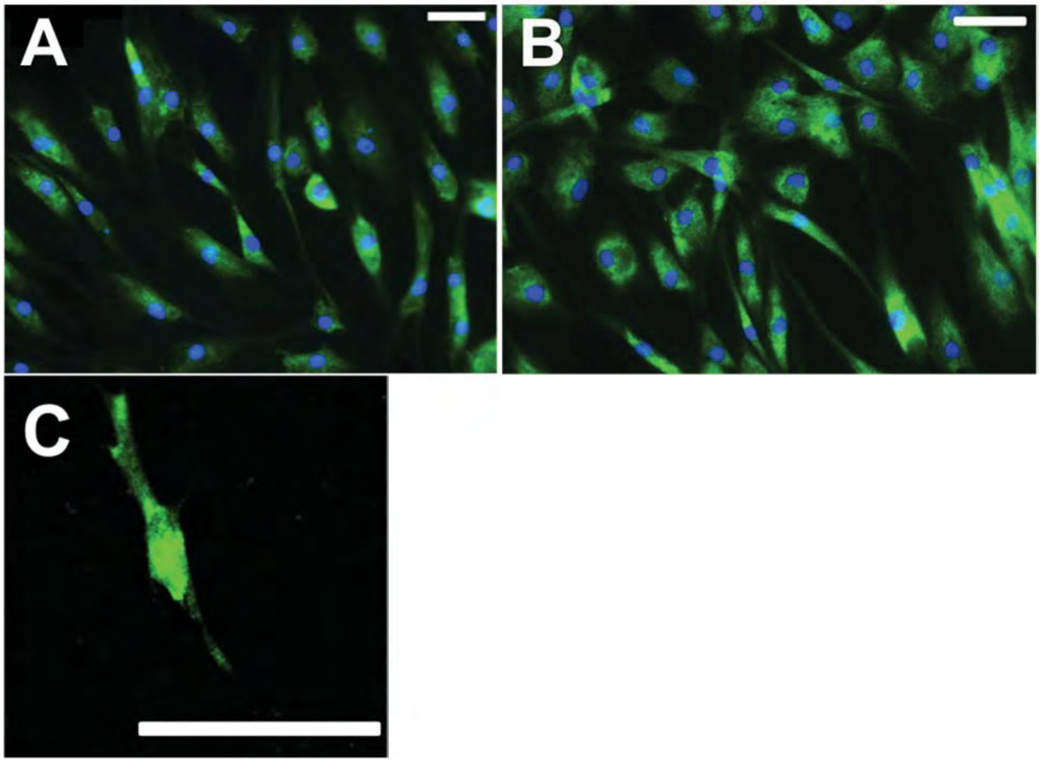
Figure 6. Pixels on a TMA were classified into fibroblast or myofibroblasts based on their average spectra as shown here. Overall normalized absorption was higher for the fibroblast class compared with myofibroblasts (A). However, in the C-H stretching region (B), myofibroblasts show stronger absorption compared with fibroblasts in the three peaks noted.

Figures

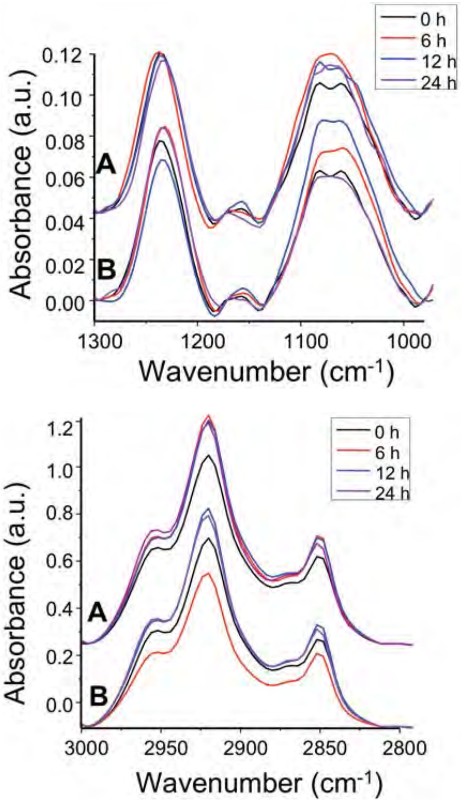
1


A
epithelial
fibroblast
B
fibroblast
epithelial
collagen
fibroblast
epithelial
collagen
cell-type specific analysis

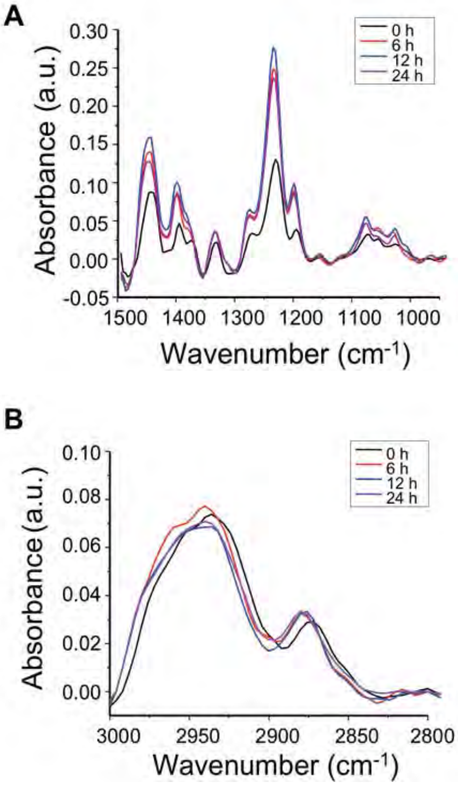
2



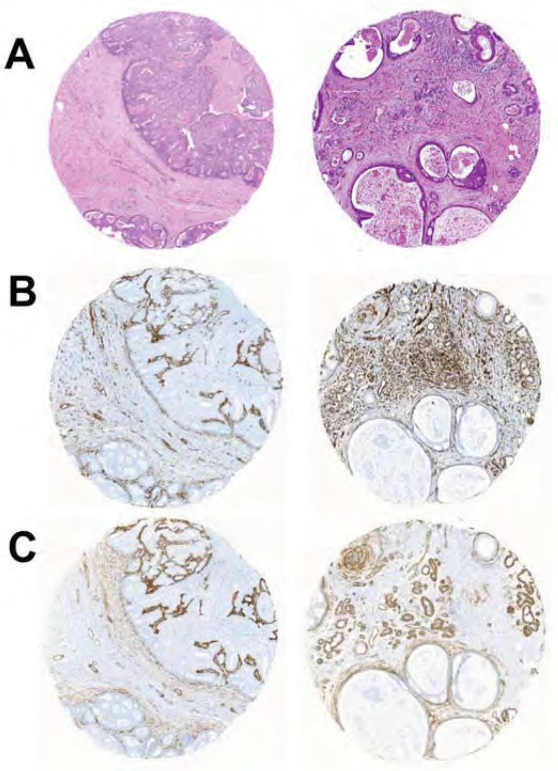
3


0 h
6 h
12 h
24 h
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Absorbance (a.u.)
A
B
1300
1200
1100
1000
Wavenumber (cm-1)
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Absorbance (a.u.)
A
B
0 h
6 h
12 h
24 h
3000
2950
2900
2850
2800
Wavenumber (cm-1)

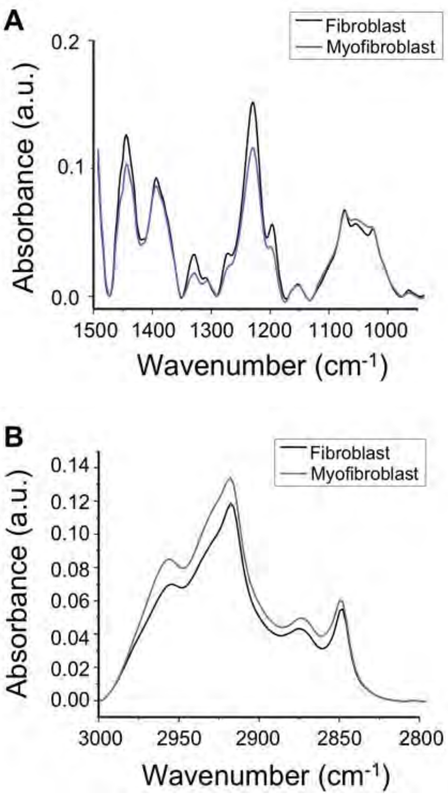
4


A
0.30
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
Absorbance (a.u.)
1500
1400
1300
1200
1100
1000
Wavenumber (cm^{-1})
0 h
6 h
12 h
24 h
B
0.10
0.08
0.06
0.04
0.02
0.00
Absorbance (a.u.)
3000
2950
2900
2850
2800
Wavenumber (cm^{-1})
0 h
6 h
12 h
24 h

5



6


A
Fibroblast
Myofibroblast
Absorbance (a.u.)
0.2
0.1
0.0
1500
1400
1300
1200
1100
1000
Wavenumber (cm-1)
B
Fibroblast
Myofibroblast
Absorbance (a.u.)
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
3000
2950
2900
2850
2800
Wavenumber (cm-1)